Structural and Molecular Dynamics Simulation Studies Support Symplekin's Protein Scaffolding Role
and
A Novel Fold in the TraI Relaxase-Helicase C-Terminus is Essential for Conjugative DNA Transfer

Sarah A. Kennedy

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Department of Chemistry (Biological Chemistry).

Chapel Hill
2009

Approved by:
Matthew Redinbo, Ph.D.
Linda Spremulli, Ph.D.
Gary Pielak, Ph.D.
Dorothy Erie, Ph.D.
Kevin Slep, Ph.D.

ABSTRACT

Sarah A. Kennedy
Structural and Molecular Dynamics Simulation Studies Support Symplekin's Protein Scaffolding Role and A Novel Fold in the TraI Relaxase-Helicase C-Terminus is Essential for Conjugative DNA Transfer
(Under the direction of Professor Matthew R. Redinbo)

The majority of eukaryotic pre-mRNAs are processed by 3'-end cleavage and polyadenylation, although in metazoa the replication-dependant histone mRNAs are subject only to 3'-end cleavage and not polyadenylation. The macromolecular complex responsible for processing both canonical and histone pre-mRNAs contains the ~1,160-residue protein Symplekin. Secondary structural prediction algorithms identified putative HEAT domains in the 300 N-terminal residues of all Symplekins of known sequence. The structure and dynamics of this domain was investigated to begin to elucidate the role Symplekin plays in mRNA maturation. The crystal structure the *D. melanogaster* Symplekin HEAT domain was determined to 2.4 Å resolution using SAD phasing methods. The structure exhibits five canonical HEAT repeats along with an extended 29 amino acid loop between the fourth and fifth repeat (loop 8) that is both unique and conserved in Symplekin sequences. Molecular dynamics simulations of this domain show that loop 8 dampens the overall motion of the HEAT domain, therefore providing a stable surface for potential protein-protein interactions. HEAT domains are often employed for such macromolecular contacts, and the Symplekin HEAT region structurally aligns with several established scaffolding proteins. Taken together, these

data support the conclusion that the Symplekin HEAT domain serves as a scaffold for protein-protein interactions essential to the mRNA maturation process.

The TraI relaxase-helicase is the central catalytic component of the multi-protein relaxosome complex responsible for conjugative DNA transfer (CDT) between bacterial cells. CDT is a primary mechanism for the lateral propagation of microbial genetic material, including the spread of antibiotic resistance genes. The 2.4 Å resolution crystal structure of the C-terminal domain of the multifunctional *E. coli* F plasmid TraI protein (TraI-CT) is presented, and specific structural regions essential for CDT are identified. The crystal structure reveals a novel fold composed of a 39-residue N-terminal α-helical extension connected by a proline-rich loop to a compact α/β core domain. Both the globular nature of the α/β-domain and the presence and rigidity of the proline-rich loop are required for DNA transfer and single-stranded DNA binding. Taken together, these data establish the specific structural features of this non-catalytic domain that are essential to DNA conjugation.

DEDICATION

To my husband Chris and my parents Ron and Sandy Huffman, I could not have accomplished this without your continued encouragement and support.

ACKNOWLEDGEMENTS

Numerous people have walked with me during my graduate school journey. First, I would like to thank Dr. Matthew Redinbo for his guidance and financial support during my time in his laboratory. Matt allowed me to pursue areas of science that I found particularly interesting, even though they were outside of his laboratories main focus. His latitude pushed me to become an independent scientist. I would also like to thank my committee members for their time and support over the past five years. Linda, Dorothy, Gary, and Kevin have all been wonderful role models and provided kind words of advice and encouragement. They are all brilliant scientists and I have enjoyed all of our discussions. Thanks also to Bob Duronio, Bill Marzluff, and Michael O'Rand, who opened their laboratories for collaborations with me. Their laboratories are filled with knowledgeable and accommodating scientists. I appreciate the time Laurie Betts and Brenda Temple spent with me to discuss technical and personal topics; they were both very instrumental in my scientific development.

Next, I would like to thank all of my fellow Redinbo lab mates, past and present. Eric Ortlund mentored me from the beginning of my career at UNC and his guidance was invaluable. Laura Guogas, my cubby mate and co-author, also played a tremendous role in my graduate school career as a friend and mentor. I think so highly of Eric and Laura and thank them from the bottom of my heart for their support and encouragement. Thanks to Jill Orans who helped me tremendously with solving my structure and thanks to Monica Frazier for help with the molecular dynamics studies. I would also like to

thank Scott Lujan, Chris Fleming, and Denise Teotico. All of these folks made life in the Redinbo Lab tons of fun and I appreciate the time they spent teaching me. To the future Redinbo alumni, Rebekah Potts, Andy Hemmert, Yuan Cheng, Joe Lomino, Monica Frazier, Mike Johnson, Bret Wallace, Jon Edwards and Dan Yao: thanks for the wonderful times and good luck with your experiments! I had the pleasure of working with two undergraduates, Ann Mast and Sung Taek Kim, who are amazingly bright young scientists and I was lucky to have worked with them both. I look forward to hearing about their career progression.

I appreciate all of my Torlon family at Solvay Advanced Polymers. Brian Stern and Geoff Underwood were great mentors and truly helped in my professional development. Faye Kirby always put a smile on my face. Jim Carnes still has me laughing with his silly emails. I'd like to thank them all for such a great experience in industry.

Finally, I would like to thank a select few who have helped me maintain sanity throughout my Ph.D. pursuit. I have sincere gratitude for Lisa Charlton who has been a friend of mine for over a decade. Her support and friendship means the world to me. She and Jill kept me from going crazy in the tall, ugly, cement tower. I'd like to thank Ralph House for putting together our small literature group and for being a thoughtful, caring friend. Laura and Scott Marsh are crazy fun and I love them both! Last, but not least, I'd like to thank the Reservoir, a great place to unwind. "Time to get to my real job…"

PREFACE

After graduating with a B.S. in Chemistry from Westminster College and working as a Research Technician at Solvay Advanced Polymers, I realized that I wanted to have a higher impact role in the chemical industry, so I knew I had to further my education. I truly enjoyed learning about biochemistry and I wanted to transition from synthetic polymer chemistry to studying human polymers: DNA, RNA and proteins. Wondering about the atomic level interactions among these polymers, I decided to pursue structural biology, specifically X-ray crystallography, so that I could understand how to solve and evaluate the three dimensional structures of biological molecules.

Matthew Redinbo, a highly regarded scientist in the field of X-ray crystallography, interviewed me for graduate school and we communicated well from the very beginning of my summer rotation. Upon joining his laboratory, he pitched numerous different options for projects, and I worked on several before deciding that I would like to bring something brand new into his laboratory. I spent about a month reading current literature and spoke to several groups on campus about potential crystallography projects. Finally, we set up collaborations with Dr. William Marzluff and Dr. Bob Duronio of the UNC Biology Department to study proteins involved in messenger RNA processing. The work characterizing Symplekin, a protein central to mRNA processing, is addressed in the first five chapters of this manuscript.

Besides working on Symplekin, I spent the past year studying TraI, a protein that has been a central focus in the Redinbo Laboratory. The sixth chapter is focused on

solving the structure of the previously uncharacterized TraI C-terminal domain and determining its role in conjugative DNA transfer. The final chapter details preliminary work to characterize the central GTPase domain of TraI, including structural predictions through sequence analysis and preliminary assay development. A summary of my short-term work on the expression and purification of human spermatozoa protein, Eppin, is also discussed in the final chapter.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

None

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF ABBREVIATIONS AND SYMBOLS

Å	angstrom
apf	atomic position fluctuations
ATP	adenosine triphosphate
AU	asymmetric unit
BM	bending magnet
BME	β-mercaptoethanol
CDT	conjugative DNA transfer
CPSF	cleavage and polyadenylation specificity factor (multi-protein complex)
CstF	cleavage stimulation factor (multi-protein complex)
C-term	C-terminal region of protein
Da	Dalton
°	degree
Δ	delta (change)
DLS	dynamic light scattering
Dm	*Drosophila melanogaster*
DNA	deoxyribonucleic acid
DSE	downstream element (usually referring to GU-rich sequence)
DTT	dithiothreitol
EM	electron microscopy
FP	fluorescence polarization
fs	femto second ($1x10^{-12}$ second)
GST	glutathione S-transferase

HEAT	helical repeat named after proteins **H**untingtin, **E**longation factor 3, PR65/**A**, and kinase **T**OR
hi	human intestine
6xhis	six histidine residues
I	Intensity
IPTG	isopropyl beta-D-thiogalactopyranoside
kD	kilodalton
LDH	lactate dehydrogenase
LB	laria broth
LRH-1	liver receptor homologue 1
M	molarity
MAD	multi-wavelength anomalous dispersion
MD	molecular dynamics
MDCK	Madin-Darby canine kidney epithelial cells
MDS	molecular dynamics simulation
MBP	maltose-binding protein
mg	milligram
mM	millimolar
mL	milliliter
MR	molecular replacement
mRNA	messenger ribonucleic acid
Ni	nickel
nM	nanomolar
nm	nanometer

ns	nanosecond
OD	optical density
PDB	Protein Data Bank by RCSB
PEG	polyethylene glycol
PK	protein kinase
PMSF	phenylmethanesulphonylfluoride
poly(A)	polyadenylic acid
Pta1	Symplekin homologue in yeast
RMSD	root mean squared deviation
SAD	single wavelength anomalous dispersion
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SeMet	selomethionine
SER-CAT	southeast regional collaborative access team
ss	single-stranded
T4SS	Type four secretion system
TB	terrific broth
TEV	Tobacco Etch Virus
TFSS	Type four secretion system
TnT	Translation and transcription kit from Promega
UNC-CH	University of North Carolina at Chapel Hill
UTR	untranslated region of RNA molecule
WT	wild-type

Chapter 1. Introduction to Symplekin

1.1 Introduction

In this scientific age, multidisciplinary approaches and collaborations between scientific groups have become imperative for understanding the complexities involved in even the simplest of biochemical processes. As a scientist, I understand the importance of a multi-faceted approach to answering tough questions. In the following five chapters, I describe the work I have done to understand a single protein, Symplekin, and how its structure relates to its biological function in messenger RNA processing and cellular tight junctions. My structure and dyanimics results will be complemented by two collaborating laboratories, who are also working to understand Symplekin's biochemical and genetic role in the model system *D. melanogaster*. Deirdre Tatomer, under the direction of geneticist Dr. Bob Duronio, is exploring the phenotype of a Symplekin-null fly, and also designing experiments to visualize different cellular junctions for the presence of Symplekin. Dr. Mindy Steiniger, under the direction of biochemist/cell biologist Dr. William Marzluff, is working to understand Symplekin's role in binding transcription factors and also in histone messenger RNA processing. This introductory chapter will survey the current literature to detail Symplekin's role in 3'-end mRNA processing and at cellular tight junctions. Control of Symplekin by sumoylation and phosphorylation will be discussed, followed by a discussion of Symplekin's interactions with several transcription factors.

1.2 Canonical messenger RNA 3'-end processing

The central dogma of molecular biology describes the flow of genetic information from DNA to RNA to protein; DNA is transcribed to RNA, which is translated into protein. In order for RNA to be translated to protein, several processes must occur to prepare the messenger RNA. These processes include adding both a 5'-end 7-methylguanosine cap and a 3'-end polyadenylic acid (poly(A)) tail, and splicing of non-protein coding introns. The 5' cap and the introns splicing are not studied herein. The poly(A) tail on the mRNA has multiple purposes, each of which is important to proper cellular function[1,2]. If the 3'-end of the mRNA is not properly formed, developmental problems can occur due to the instability of the message or improper levels mRNA translation[1]. Several cases of ocular dystrophy, thalassemias and lysosomal storage disorder have been linked to aberrant 3'-end polyadenylation[1]. Specifically, depletion of CstF-64, an essential 3'-end processing factor, disrupts proper 3'-end formation and leads to cell cycle arrest and apoptosis[3]. The addition of a 30-nt poly(A) tail stabilizes the mRNA transcript and protects the message from unregulated degradation, since deadenylation is one of the principle ways that mRNA degradation is triggered[4-6]. Also, the poly(A) tail promotes the export of the mRNA from the nucleus to the cytoplasm for translation initiation[7,8] and transcription termination is promoted by the synthesis of the poly(A) tail[9]. Studying the proteins, RNA elements and mechanism of 3'-end cleavage and polyadenylation will lead to a better understanding of the intricate regulatory nature of transcription termination and translation initiation.

The addition of the poly(A) tail requires two steps. First, an endonucleolytic cleavage occurs at a conserved CA sequence in the 3' untranslated region (UTR). Then, the 3'-end is subject to adenylic acid polymerization. While this seems like a fairly simple two step

process, many RNA recognition sites and proteins are necessary for proper 3'-end processing. Many recent reviews have discussed the cleavage and polyadenylation proteins and RNA recognition sequences in plant, yeast, and mammalian systems[1,2,5,10]. **Figure 1.1**, taken from Dominski *et al.*, shows the main machinery involved in canonical mRNA processing in mammals[11]. The following section discusses the main RNA and protein elements responsible for the 3' cleavage and polyadenylation reactions.

The mammalian RNA recognition sites include a U-rich sequence, followed by an AAUAAA sequence that is approximately 30 nucleotides upstream of the CA cleavage site and then a GU-rich element follows 30 nucleotides downstream of the cleavage site[1,12]. The AAUAAA sequence is necessary for both polyadenylation and cleavage; mutations in this region strongly inhibit processing[1,13,14]. While the GU-rich downstream element (DSE) sequence can be variable, position is important, and the DSE must be located within 30 nucleotides of the cleavage site[15]. Yeast and plants have both the U-rich and AAUAAA sequences; however, they are slightly less conserved than in mammals. The CA cleavage site is also slightly different: the A is conserved, but the C can be either C or U in these two species.

In mammals, four multi-protein complexes (CstF, CFI_m, $CFII_m$, and CPSF), poly(A) polymerase (PAP), poly(A) binding protein (PABP), RNA Polymerase II C-terminal domain (PolII CTD), and Symplekin are all required for 3'-end processing[12]. The mammalian CPSF (Cleavage and Polyadenylation Specificity Factor) complex includes subunits CPSF-73, CPSF-100, CPSF-30, CPSF-160 and hFip1, each of which binds to RNA. CPSF-73 is the main cleavage enzyme responsible for cleaving the CA RNA sequence in the 3' UTR[16]. The CstF complex is composed of CstF-50, -64, and -77. CstF-50 and CstF-77 bind to PolII

CTD, CstF-64 binds to the DSE in the mRNA, and CstF-77 links the CstF to the CPSF complex. CFI_m is regulated by phosphorylation and provides additional recognition of the pre-mRNA substrate, which helps to define the correct polyadenylation site and binding of PAP[17,18]. $CFII_m$ has two subunits, Pcf11 and Clp1: in yeast Pcf11 binds to PolII CTD, CstF and Symplekin, while Clp1 has an ATP-binding domain and binds to CPSF and CFI_m.

The remaining factors include PAP, PABP, PolII CTD and Symplekin. Poly(A) polymerase (PAP) is the main protein required to add AMP molecules to the pre-mRNA[19,20]. Poly(A)-binding protein (PABP) binds to stretches of 11-14 poly(A) nucleotides and regulates the proper tail length[21,22]. The CTD of PolII is necessary for 3'-cleavage in mammals but not yeast, and it has been shown that *in vitro* 3'-end processing reactions are stimulated by phosphorylated CTD[23-25]. Symplekin forms a complex with CPSF and CstF and in has been shown to be required for both cleavage and polyadenylation in yeast[12,26,27]. Symplekin is proposed to be a scaffold that is essential for forming the properly oriented 3' processing machinery.

As described, this seemingly simple two step reaction of cleavage and polyadenylation requires multiple factors and intertwined protein-protein and protein-RNA interactions. Separating functions of each protein has been very challenging because of the interdependent nature of the interactions. In fact, the most recent experiments on the human pre-mRNA 3' processing complex indicate that this purified complex contains over 85 proteins, 50 of which are important to correlate 3'-end processing with other cellular processes[28]. Further information about these interactions is outlined in several reviews[1,12,20,29]. As a structural biologist, I am interested in understanding the structural basis for the formation of the main components of the 3'-end processing machinery. At this

point, the structures of many of the proteins have been determined, including portions of CPSF 73, CPSF 100, CPSF 30, CstF 64, CstF 77, PAP-Fip1, and CF Im-25. Thus, I focused on understanding an essential protein in this process that had no structural information available to date: Symplekin.

1.3 Symplekin's role in canonical mRNA 3'-end processing

Yeast Symplekin (named Pta1) is required for both cleavage and polyadenylation of canonical mRNA[27]. In the yeast polyadenylation system, Pta1 has been shown to be present in a complex with the human homologues of CPSF-73, CPSF-100, and five other factors that have no known mammalian counterparts: Ssu72p, Swd2p, Syc1p, Pti1p, Glc7p[30-33]. Symplekin has been shown to bind with both CstF-64 and CstF-77, but in a mutually exclusive manner [26]. Thus, Symplekin is proposed to be a scaffold upon which the CstF complex is assembled in the correct orientation. Unpublished work from the Marzluff laboratory (UNC-CH) shows that Symplekin binds to CPSF-73 and CPSF-100. Symplekin is colocalized with CPSF-100 in Cajal bodies during oocytes maturation and binds directly to CPSF and CPEB to regulate protein-protein interactions in cytoplasmic polyadenylation[34,35]. Taken together, these data suggest Symplekin is a protein essential for mRNA processing by providing a scaffold on which protein-protein interactions can occur.

1.4 Symplekin's role in histone mRNA 3'-end processing

Histone mRNAs, are different than canonical mRNA in that they only require a 3'-end cleavage and not 3'-end polyadenylation. Histone mRNA 3'-end structure is that of a stem-loop sequence. **Figure 1.2** illustrates the difference between the 3'-end structure of

canonical and histone mRNAs. The factors involved in histone mRNA processing are slightly different than those required for canonical mRNA. **Figure 1.3**, taken from Dominski *et al.*, shows the macromolecular machinery involved in histone mRNA 3'-end processing[11]. Common factors to both canonical and histone mRNA processing include Symplekin, CPSF-73 and CPSF-100. Unpublished data by Sullivan, Steiniger and Marzluff indicate that CPSF-73, Symplekin and CPSF-100 form a core complex that is required for histone pre-mRNA processing in *D. melanogaster.*

Symplekin was initially discovered to play a part in histone processing by Kolev and Steitz, who identified Symplekin as the main component of the heat labile factor (HLF), which was responsible for the disruption of histone mRNA 3'-end processing in heat-treated HeLa nuclear extracts[36]. Adding Symplekin restored histone mRNA processing and demonstrated the reassembly of the multi-subunit complex containing CPSF and CstF[36]. Once again, data points towards Symplekin playing a role as a scaffold for the integrity of the 3'-end processing machinery.

1.5 Sumoylation and phosphorylation regulate Symplekin

Vethantham *et al*. discovered that sumoylation modulates the activities of both CPSF73 (the endonucleolytic cleavage factor for 3'-end mRNA processing) and Symplekin[37]. Highly conserved sumoylation sites in Symplekin and CPSF-73 are targeted by SUMO-2/3 and desumoylation of Symplekin or CPSF-73 by adding a SUMO protease reduces 3'-end processing and inhibits the proper assembly of the 3'-end processing complex[37]. In addition to sumoylation, phosphorylation also exerts a level of regulation in mRNA processing. In the yeast polyadenylation complex, He *et al.* demonstrated that

reduction of phosphatase Glc7p levels lead to a decrease in the length of poly(A) tails and also to the accumulation of phosphorylated Pta1[33]. Addition of unphosphorylated Pta1 rescued the 3'-end processing reactions. These two results show that Symplekin is controlled by two common regulatory molecules and indicates that Symplekin has a regulatory role for 3'-end processing in addition to scaffolding CPSF and CstF in the correct orientation for cleavage and polyadenylation.

1.6 Symplekin is also found at epithelial cell tight junctions

Before Symplekin's role in mRNA processing was documented, this protein was shown to be involved in cellular tight junctions of several epithelial cell lines. Tight junctions, also known as zonula occludens, composed mainly of the transmembrane protein occludin, form a gasket-like link between epithelial or endothelial cells to provide a fence-like hydrophobic barrier between lumina[38,39]. A program to provoke antibody production for tight junctional plaque proteins produced a monoclonal antibody specific to Symplekin. Using immunofluorescence, cell fractionation, RNA isolation, DNA sequencing and protein characterization methods, Keon *et al*. revealed a 1142 residue human protein present at tight junctions[40]. They named the protein Symplekin, which in Greek means "to tie together, to be intertwined, to weave". In all of the cell types analyzed, Symplekin was found in the nucleoplasm, but only in cell lines forming tight junctions was Symplekin found to be present in the plaque of zonula occludens[40,41]. Keon's study revealed that Symplekin was very specifically found in the zonula occludens, rather than zonula adherins[40]. Thus, Symplekin was able to be used as a differential marker for identification of cells forming zonula occludens, which has been proposed to be useful in assessing tight junctional

formation in certain carcinomas[39,40]. Another study shows the presence of Symplekin in the outer limiting zone of the retina, a new type of adhering junction[42]. The exact purpose for Symplekin at these tight junctions is not known. Presently, no studies have linked Symplekin's role in tight junctions to its role in mRNA processing. However, it is possible that Symplekin plays a scaffolding role at tight junctions, just as it is predicted to scaffold factors for mRNA processing.

1.7 Symplekin modulates transcription factors ZONAB and HSF1

Symplekin also serves as bridging factor between the 3'-end machinery and transcription regulators. Specific examples of Symplekin's role in transcription regulation by interaction with transcription factors HSF1 and ZONAB have been shown[43,44]. HSF1 binds to the *HSP* gene promoters to upregulate expression of heat shock proteins in response to cellular stress. The N-terminal 125 residues of Symplekin interact with heat shock factor 1 (HSF1) in the nuclei of heat stressed cells[44]. Disruption of the HSF1-Symplekin interaction reduced the polyadenylation of mRNA encoding HSP70 thus limiting the ability of the cell to respond to stress[44]. CstF64 was also found co-localized with the HSF1-Symplekin complex. Thus, Symplekin's role in helping to upregulate HSP70 expression during heat stress has been established and demonstrates Symplekin's role in coupling mRNA polyadenylation to protein expression.

ZONAB is found in the nucleus, where it participates in transcription regulation, and is also found at epithelial cell tight junctions bound to ZO-1. Kavanagh *et al.* demonstrate through confocal microscopy and epifluorescence that Symplekin and ZONAB are colocalized in the nucleus of MDCK (canine epithelial cells) and this nuclear interaction can

be immunoprecipitated and also reconstituted with recombinant proteins[43]. Through a luciferase reporter assay, Symplekin was observed to modulate ZONAB's transcriptional activity[43]. Undoubtedly, future studies will probe more deeply into the mechanism that Symplekin uses for regulating transcription factors and show more links between transcription regulation and mRNA processing. Symplekin's role in 3'-end messenger RNA processing has clear implications for it's regulation of available mature RNA molecules that can be translated to protein. Symplekin interacts with transcription factors like HSF1 and ZONAB and may bridge the 3'-end processing machinery to the transcription regulation machinery.

1.8 Understanding Symplekin through crystallography and molecular dynamics

Table 1.1 summarizes the published Symplekin protein-protein interactions. The table includes species, specific residues if known, and method of detection. Most of these interactions have not been studied in sufficient detail to pinpoint the exact regions of binding within each protein. However, three independent studies, in yeast or human cells, have shown the N-terminus of Symplekin binds to transcription factor HSF1 and phosphatases Ssu72 and Glc7p[33,44,45]. Thus, this section of Symplekin appears to be a protein scaffold important for regulation of mRNA processing by phosphorylation and transcription factor modulation. I used a combination of crystallography and molecular dynamics simulations to investigate the N-terminal region of Symplekin. The following chapters discuss the identified domains within Symplekin, the structure determination of the N-terminal HEAT domain and the role that a conserved unique loop plays in the dynamics of this domain. Throughout these experiments, it becomes increasingly clear that Symplekin's HEAT

domain does in fact form a structural scaffold with the ability to accommodate multiple diverse binding partners.

1.9 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 1.

Figure 1.1 Canonical mRNA 3'-end processing machinery bound to an mRNA molecule

Figure 1.2 Cartoon representations illustrating the difference in 3'-end processing of canonical and histone mRNAs

Figure 1.3 Histone mRNA 3'-end processing machinery bound to histone mRNA

Table 1.1 Symplekin's known interactions and methods of detection

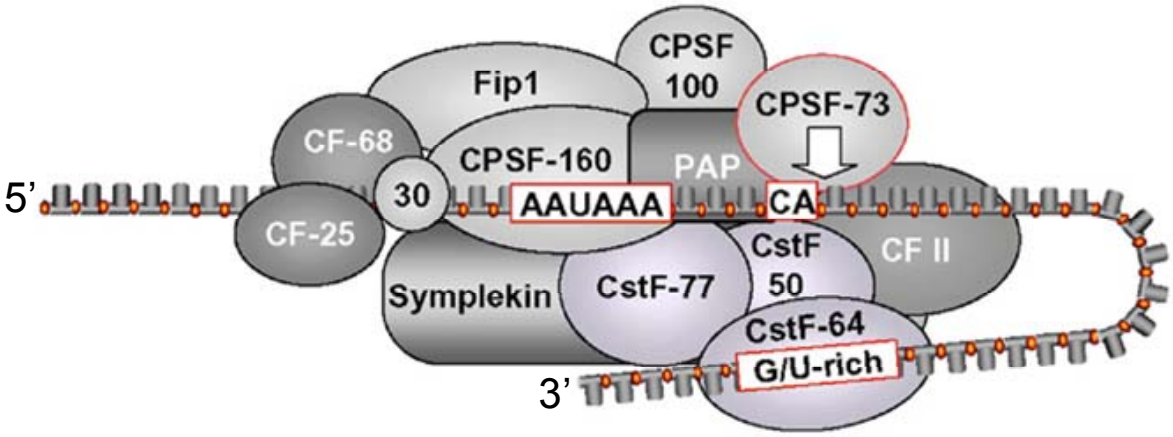


Figure 1.1 Canonical mRNA 3'-end processing machinery bound to an mRNA molecule

Figure legend from Dominski *et al.*[11] The processing machinery involved in cleavage/ polyadenylation. The arrow indicates the site of cleavage by CPSF-73. The position of some of the components in the complex has not been experimentally supported. The model does not include the CTD that is required for *in vitro* cleavage/polyadenylation under certain experimental conditions[11].

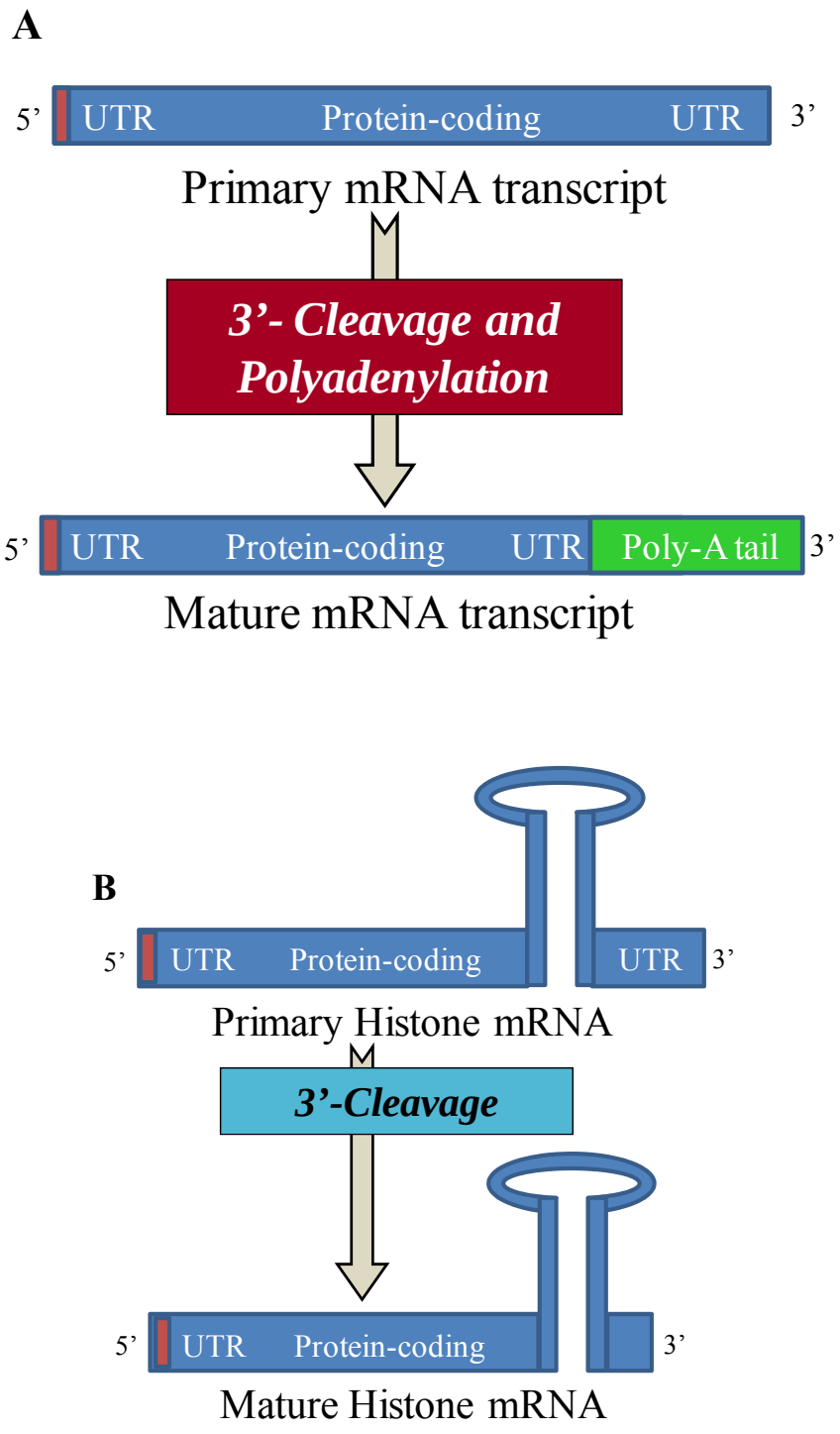


Figure 1.2 Cartoon representations illustrating the difference in 3’-end processing of canonical and histone mRNAs

Illustration of 3’-end processing reactions for (**A**) canonical mRNA and (**B**) histone mRNAs. UTR stands for untranslated region.

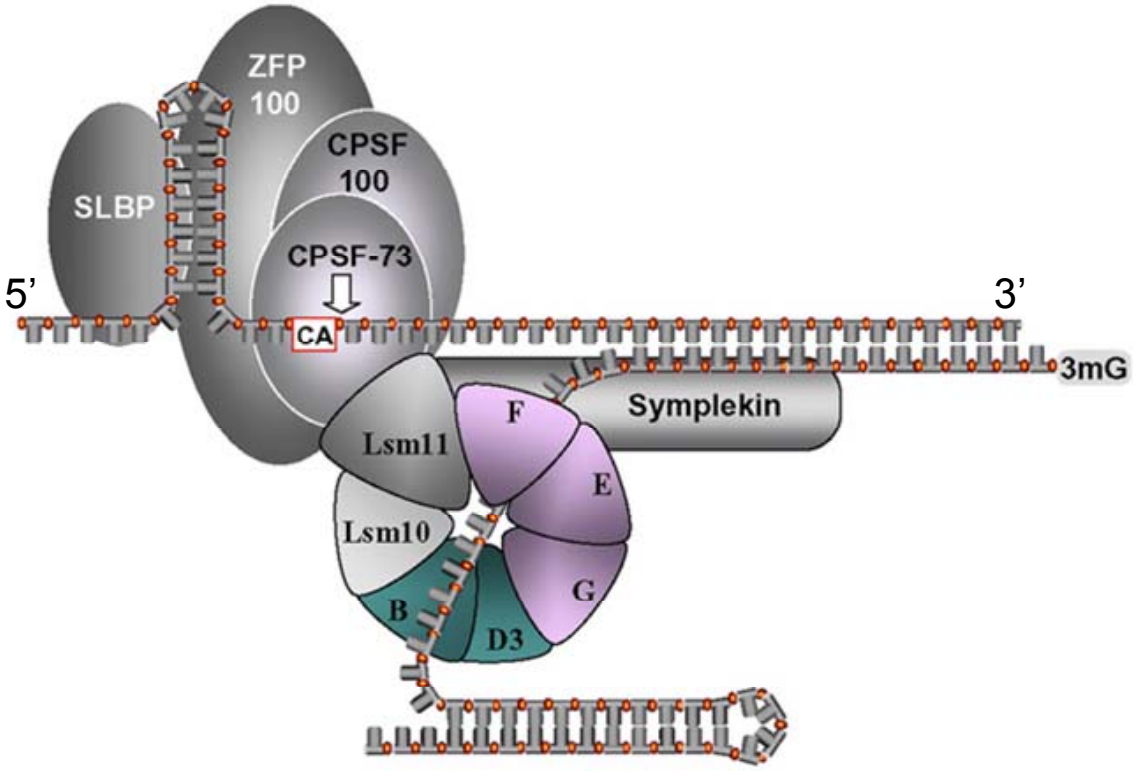


Figure 1.3 Histone mRNA 3'-end processing machinery bound to histone mRNA[11]

Figure legend from Dominski *et al.* Known components of the 3′ end processing machinery cleaving histone pre-mRNAs. The position of Symplekin in the complex is hypothetical and the presence of CPSF-100 has not been supported experimentally[11].

SLBP (stem-loop binding protein) and ZPF100 recognize the mRNA stem loop. CPSF73 is the main endonucleolytic enzyme and is thought to dimerize with CPSF 100. Proteins B, D3, G, E, F, Lsm11 and Lsm10 are Sm core proteins, which are part of the U7 snRNP responsible for recognizing the 3' histone downstream element that is distant from the CA cleavage site.

Table 1.1 Symplekin's known interactions and methods of detection

Other protein	*in vivo*	*in vitro*	Detection method(s)	Species/cell line	Other details	Citation
occludin	x		immunofluorescence colocalization	hiHT-29	colocalized at tight junction	Kavanagh
ZONAB/DbpA	x		immunofluorescence colocalization	canine, MDCK	colocalized at nucleus and tight junction	Kavanagh
	x		coimmunoprecipitation	hiCaco-2, MDCK	ZO-1 precipitates also	Kavanagh
		x	pull-down with recombinant protein	-	-	Kavanagh
	x		luciferase reporter transcription assay	MDCK, hiHT-29	additive transcription repression	Kavanagh
ZO-1	x		coimmunoprecipitation	hiHT-29	-	Kavanagh
HSF1	x	x	Yeast two hybrid and pull down assays	mouse	mSymplekin 1-124 interacts with HSF1	Xing
	x		immunoprecipitation	human	nuclear extracts of heat shocked HeLa cells	Xing
	x		immunofluorescence colocalization	human	nuclear extracts of heat shocked HeLa cells	Xing
CstF-64	x		immunoprecipitation, indirect thru HSF1	human	nuclear extracts of heat shocked HeLa cells	Xing
	x		coimmunoprecipitation	human	HeLa cell nuclear extracts, CstF-64 100-200	Takagaki
		x	coimmunoprecipitation	human	in vitro tranlsated proteins	Takagaki
CstF-77		x	coimmunoprecipitation	human	in vitro tranlsated proteins	Takagaki
Ssu72	x	x	immunoprecipitation, pull down assay	yeast	-	He
Sub1	x	x	immunoprecipitation, pull down assay	yeast	-	He
Swd2p		s	pull down assay	yeast		Dichtl
Cft2p/Ydh1p		x	pull down assay	yeast	Cft2p/Ydh1p is a human CPSF-100 homologue	Kyburz
brr5	x		coimmunoprecipitation	yeast	interaction observed in purified CFII	Zhao
	x	x	Yeast two hybrid and pull down assays	yeast	brr5 is a human CPSF-73 homologue	Zhelkovsky
Syc1p		x	pull down assay	yeast		Zhelkovsky
Glc7		x	pull down assay	yeast	Glc7 phoshpatase binds 101-201 of pta1	He
CPEB	x		coimmunoprecipitation	xenopus oocytes	-	Barnard
CPSF-100	x		coimmunoprecipitation	xenopus oocytes	-	Barnard

Chapter 2. Cloning, expressing and crystallizing the HEAT domain of *Drosophila melanogaster* Symplekin

2.1 Introduction

As described in chapter one, Symplekin is involved in transcriptional regulation and mRNA 3'-end processing, it is present at tight junctions and is regulated by sumoylation and phosphorylation. The diverse functionality and the lack of structural information made Symplekin a fascinating target for biophysical and structural characterization. As mentioned previously, collaborators at UNC-CH are also working on Symplekin, specifically in the *D. melanogaster* model system. I chose to work with this species of Symplekin, so that our work would directly correlate. This chapter focuses on the work leading up to solving the crystal structure of the Symplekin HEAT domain. First, discovery of well-folded structural motifs by primary sequence analysis is discussed. Then, cloning and test expression procedures for various Symplekin constructs are outlined. Finally, the purification, crystallization, and structure determination of the N-terminal Symplekin HEAT domain is described.

2.2 Structural prediction of Symplekin domains

Generally, the first step in designing a crystallization target is to examine the primary amino acid sequence for predicted secondary structure and domains. In general, large regions that are predicted to be disordered can disrupt crystal packing and are also more

prone to proteolysis during protein purification. The following software programs were utilized to find symplekin orthologs and predict the structural elements within Symplekin: BLAST[46], Jpred[47], PHYRE[48], pFam[49], InterProScan[50], ScanSite[51], PredictProtein[52], RONN[53] and COILS[54]. BLAST identifies homologous proteins to that of your search sequence, whereas pFam, InterProScan, PredictProtein and ScanSite search your sequence against classes of characterized functional domains. Jpred and PHYRE assign secondary structural elements to your primary sequence. PHYRE also creates models of your protein based on similar published structures and scores them with E-values. RONN specifically finds areas of predicted disorder and COILS predicts regions of coiled-coils within the primary sequence.

Domain predictions for full length Symplekin were made by combining the predictions from each of these sites (**Figure 2.1**). The disordered regions include 1-18, 452-544, and 1116-1165. A HEAT domain was predicted between residues 19-271, a region with predicted alpha helical structure follows immediately after the HEAT domain at residues 272-451, and the C-terminal portion between 545-1116 is predicted to fold into an armadillo domain. No additional domains were predicted for Symplekin, thus a catalytic activity is most likely not going to be attributed to Symplekin. Rather a scaffolding role for protein or nucleic acid interactions is its predicted function based on its structural elements.

2.3 Construct design, cloning and test expressions

The original full length cDNA encoding *D. melanogaster* Symplekin in the pGEX vector was obtained from Dr. Mindy Steiniger in Dr. Bill Marzluff's laboratory at UNC-CH. I cloned the putative domains into the pMCGS9 vector with either 6xHis or 6xHis-MBP tags

utilizing ligation independent cloning[55]. **Table 2.1** lists all of the constructs that I cloned and test expressed in *E. coli*. **Figure 2.2** shows a representative test expression for Symplekin 634-1082 and 669-1082 in the pMCGS9 plasmid (from Sondek lab at UNC-CH). In this case, two different amounts of IPTG were added to induce protein expression under the lac operon. Unfortunately, this protein was present in the insoluble fraction. No attempt at expressing the C-terminal armadillo domain in bacterial cells resulted in soluble protein expression. Full length Symplekin protein did express in bacterial cells, however, upon purification, there was always a cleavage event, which cut the protein roughly in half. Despite attempts to stabilize the full length protein with the use of protease inhibitors, Triton X-100, salt concentrations varied from 0-500 mM and a pH range of 7-8, the full length protein never was able to resist truncation. Thus, crystallization quality full length protein was not obtained. The result of all of the test expressions was the soluble overproduction of Symplekin 19-271 in the LIC-MBP vector in BL21 cells. Since HEAT domains are common in scaffolding proteins, three regulatory proteins bind to this region of Symplekin, and no structural information had yet been published on Symplekin or its homologues, I began to focus on crystallizing this HEAT domain to understand its' potential scaffold role.

2.4 Expression and purification of the Symplekin HEAT domain

As mentioned above, residues 19-271 of *D. melanogaster* Symplekin were cloned into the expression vector pMCGS9 (also known as the LIC plasmid), which provided N-terminal 6-histidine and maltose-binding protein (MBP) tags followed by a Tobacco Etch Virus (TEV) protease site[56]. *E. coli* BL21 (DE3) gold cells (Stratagene) were transformed with the plasmid and cells were grown at 37 °C in 1.5 L of terrific broth supplemented with

50 mg/L ampicillin until an A_{600}=1.0-1.2 at which time the temperature was dropped to 18 °C and 0.1 mM of IPTG was added to induce protein expression overnight until a final OD A_{600}=4.5. The cells were harvested by 20 min 1600 g centrifugation and resuspended in 20 mL of nickel buffer A (5 mM imidazole, 50 mM potassium phosphate, pH 7.4, 150 mM NaCl, 1 mM dithiothreitol, 0.01% sodium azide) and stored at -80°C. The cells were lysed by sonication in the presence of DNase and a protease inhibitor cocktail and then centrifuged at 40,000 g for 60 minutes to produce a cleared lysate. The protein was purified away from the lysate by nickel affinity chromatography (**Figure 2.3A**). Nickel buffer B (500 mM imidazole, 50 mM potassium phosphate, pH 7.4, 150 mM NaCl, 1 mM DTT, 0.01% sodium azide) was used to elute the protein with a gradient of 5-100% B. To cleave the 6xHis-MBP-Symplekin fusion, 2% TEV protease by mass TEV/mass Symplekin was added (**Figure 2.3B**). Protein was dialyzed into nickel buffer A during TEV cleavage. A second nickel column or amylose column purified the now un-tagged Symplekin from the 6×His-MBP tag (**Figure 2.3C**). A polishing step of size exclusion chromatography (Column: Superdex 75, GE Healthcare; sizing buffer: 10 mM HEPES, pH 8.0, 50 mM NaCl, 1 mM DTT and 0.01% sodium azide) cleaned the protein to greater than 95% purity, by SDS-PAGE (**Figure 2.3D**).

A selenomethionine-derivative of residues 19-271 of *D. melanogaster* Symplekin was grown using *E. coli* B834 cells, a methionine auxotroph cell line. Cells were grown in selenomethionine specific media (Athena) supplemented with 50 mg/L selenomethionine. The culture was grown at 37 °C until an OD of 0.6, then the temperature was reduced to 18°C and 0.1 mM IPTG was added to induce overexpression of the derivative protein overnight. Purification procedures were identical to those listed above for the native protein.

2.5 Crystallization of Symplekin HEAT domain

Typically protein crystallization begins with commercially available matrix screens that explore multiple pH values, salt conditions and precipitant concentrations. However, since the structure of Symplekin was proposed to be a HEAT domain, a literature search of similar structures was performed to develop a reasonable set of crystallization conditions. Ann Mast, an undergraduate working with my on this project, helped to put together this information and design screens around the published conditions used to crystallize other HEAT domains.

Native and selenomethionine derivative Symplekin proteins were concentrated to 3-6 mg/mL in the sizing buffer. Crystallization was performed by hanging drop diffusion at 22°C with mother liquor consisting of 0.4-0.5 M sodium citrate, 25-28% PEG 3350, 10 mM HEPES, pH 8.0, 0.01% N_3Na and 1 mM DTT. Each crystallization drop contained 1 μL of protein and 1 μL of well solution. Diamond shaped crystals grew within one week; the largest ones had the dimensions 300 μm x 60 μm x 60 μm. **Figure 2.4** shows representative Symplekin 19-271 crystals. Multiple cyro-protectant solutions similar to the mother liquor were screened by cooling them in liquid nitrogen and visualizing their transparency. Finally, crystals were cryoprotected in mother liquor plus 35% PEG 3350 or perfluoroether oil and flash-cooled in liquid nitrogen.

A fluorescence scan was performed on the selenomethionine-derived Symplekin crystal to confirm the presence of a strong selenium signal (**Figure 2.5**). Diffraction data were collected at 100K using Sector 22-BM (SER-CAT) of the Advanced Photon Source at Argonne National Laboratories. A strategy of 1° oscillation with two second exposure and collection of 360° of data was most ideal for completeness and redundancy. **Figure 2.6**

illustrates a diffraction image from the native data collection. A SAD data set was collected on the seleno-methionine-derivative Symplekin crystals at selenium edge of 0.97190 Å; a native data set was collected at 0.97958 Å. DENZO and SCALEPACK in HKL-2000 were used to index and scale the data[57]. The crystals were in the space group $P4_12_12$ with unit cell dimensions of a, b = 68.7 Å , c = 138.5 Å and α, β, γ = 90°. More detailed data collection statistics are listed in **Table 2.2.**

2.6 Data processing and structure refinement

The SGXPRO software package[58], an interface for programs including SHELXD[59] and SOLVE/RESOLVE[60], was used to solve the heavy atom sites and initial phases. A Matthews coefficient value of 2.9 indicated with 99% probability that 1 molecule was expected in the asymmetric unit with 57.6% solvent. Six methionine residues were present in Symplekin 19-271, thus six Se sites were expected. SHELXD and SOLVE identified all six Se atom positions, and the initial phase was calculated to 2.9 Å. RESOLVE was used for density modification and to provide an initial model. After these steps, the overall figure of merit was 0.69. **Figure 2.7** illustrates the *original* experimental electron density with the final model of the Symplekin protein, thus validating the phase solution found using SGXPRO[58].

The model was built by hand using COOT[61]. Initially, all helices were built with alanine residues. Loops were added over several rounds of refinement to connect the helices. Finally, side chains were placed in the model. This 2.9 Å model from the SAD data was refined using REFMAC5 at this stage to R and R_{free} values of 0.353 and 0.419, respectively. To phase the 2.4 Å native data set, the selenium-derivative model was used in molecular replacement[62]. Further refinement was conducted by building and validating the model in

coot, and employing both CNS and REFMAC5. Final statistics (**Table 2.3**) include R and R_{free} values of 0.2068 and 0.2653, respectively. For both the original SAD data and the final native data, 5% of the data were set aside for the free-R and not used at any stage of refinement. The final model, consisting of 248 residues (no density was present for residues 19-21 and 271) and 142 water molecules, was validated with PROCHECK and Molprobity[63]. **Figure 2.8** illustrates a portion of the final model in the electron density from the 2.9 Å data set (**A**) and the 2.4 Å data set (**B**). The improvement in resolution can be clearly seen in this example of electron density.

2.7 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 2.

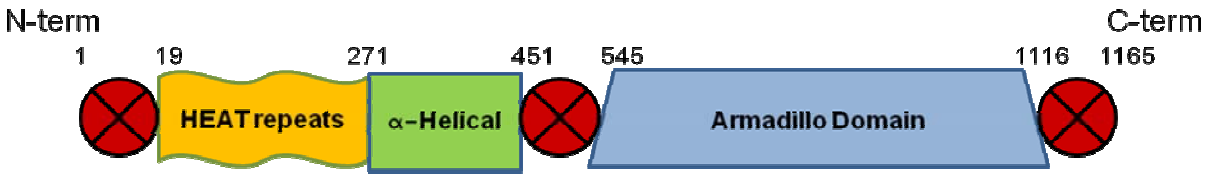


Figure 2.1 Predicted domains within *D. melanogaster* Symplekin.

The numbers above the figure represent amino acid positions. Red circles denote areas with predicted disorder. HEAT repeats are predicted between 19-271 followed by an α-helical region between 272-451. An armadillo domain is predicted for the C-terminal half of the protein. The N and C termini are labeled.

Table 2.1 List of Symplekin constructs and test expression conditions

Construct	Tags	*E. coli* Cell Line	Temperatures (°C)	IPTG (mM)
1-1165	His, MBP, GST	BL21, BL21-RIPL, BL21-RIL, BL21-Origami	37, 18	0.1, 1
1-477	His, MBP	BL21, BL21-RIPL, BL21-RIL, BL21-Origami	18	0.1, 1
19-271	His, MBP	BL21, BL21-RIPL, BL21-RIL, BL21-Origami	37, 18	0.1, 0.5, 1
528-1165	His, MBP	BL21, BL21-RIPL, BL21-RIL, BL21-Origami	18	0.1, 1
634-1082	HIS	BL21	18	0.1, 1
669-1082	HIS	BL21	18	0.1, 1

His is a 6xhistidine tag , MBP is maltose binding protein, GST is glutathione-S-transferase. All tags are on the N-terminus of the Symplekin construct.

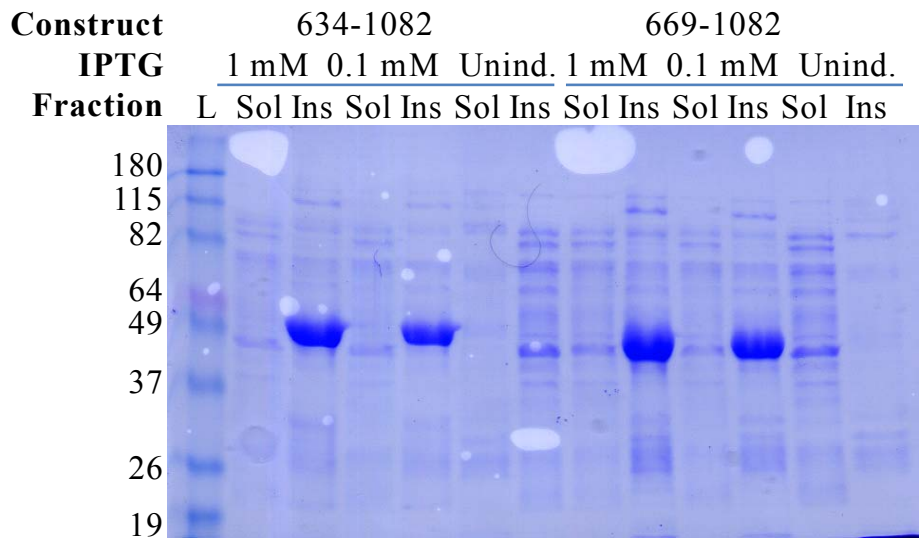


Figure 2.2 Test expression of Symplekin 634-1082 and 669-1082

SDS-page gel of the *E. coli* BL21 test expression of Symplekin constructs 634-1082 and 669-1082 in the pMCGS9 plasmid. The IPTG row represents the mM of IPTG used to induce protein expression (Unind corresponds to a sample taken before induction). The fraction row designates L for molecular weight ladder, Sol for soluble fraction and Ins for insoluble fraction. The theoretical molecular weight for 634-1082 is 52.3 kD, while 669-1082 is 48.2 kD. This is a representative gel from all of the test expressions. Samples were prepared by lysing a cell pellet from 1 mL of induced bacterial cells, then separating soluble and insoluble fractions with centrifucation at 17,000 rpm for 30 minutes. These proteins were not subject to any purification steps.

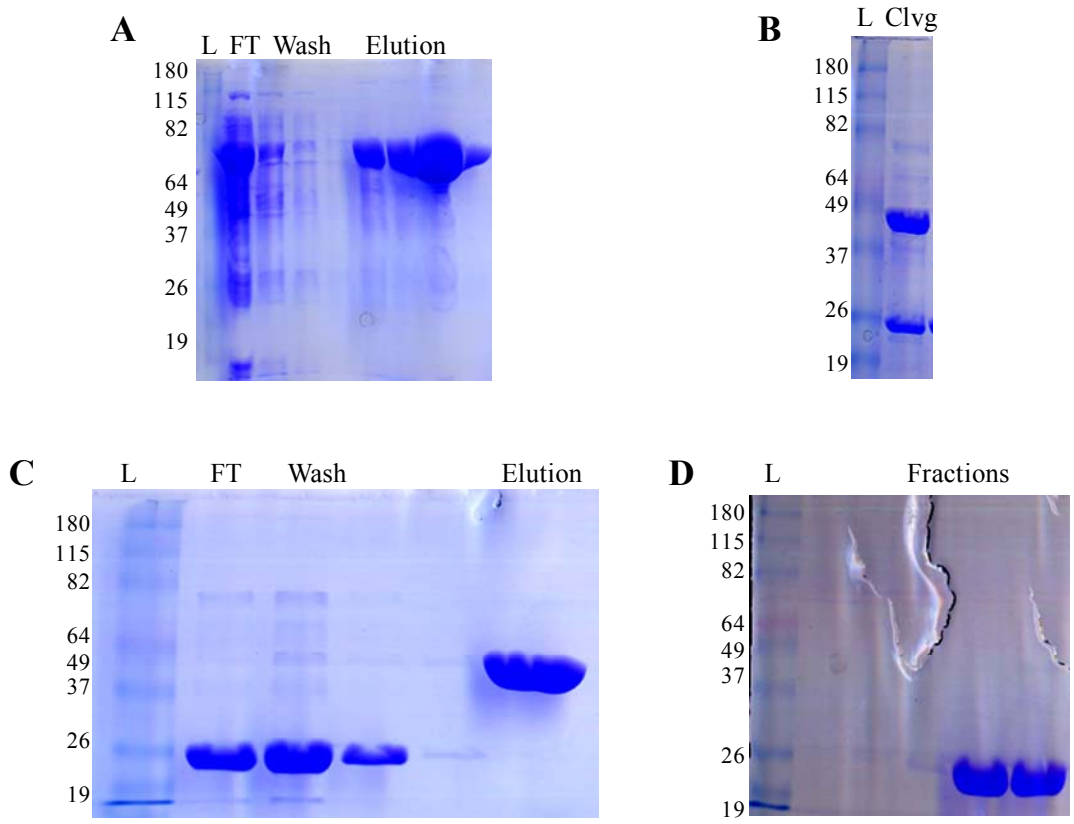


Figure 2.3 Purification of Symplekin 19-271

SDS-page gels of the purification scheme for Symplekin 19-271. L represents the molecular weight markers, FT stands for flow through, Clvg stands for a sample showing the cleavage by TEV protease. **(A)** Representative fractions from a nickel column with cleared lysate. The MBP-Symplekin protein runs at 70 kD. **(B)** Gel showing the result of the overnight TEV cleavage reaction to cleave the 42 kD MBP tag from the 28 kD Symplekin construct. **(C)** Amylose column run after the TEV cleavage reaction. Untagged Symplekin (28 kD) now flows off of the column, while MBP (42 kD) binds to the amylose resin. (A gel run on a nickel column sample at this stage has the same elution profile). **(D)** Final purification step of size exclusion chromatography yields > 95% pure untagged Symplekin 19-271.

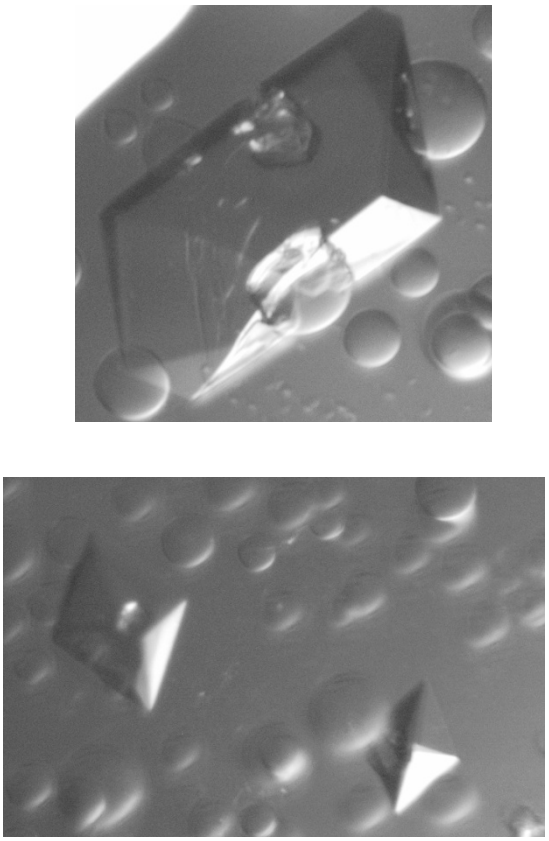


Figure 2.4 Crystals of Symplekin 19-271

This is an illustration of Symplekin crystals grown in 425 mM sodium citrate, 26% PEG 3350, 10 mM HEPES-KOH, pH 8.0, 1 mM DTT. Crystals were grown by hanging drop diffusion at 22°C after 5 days.

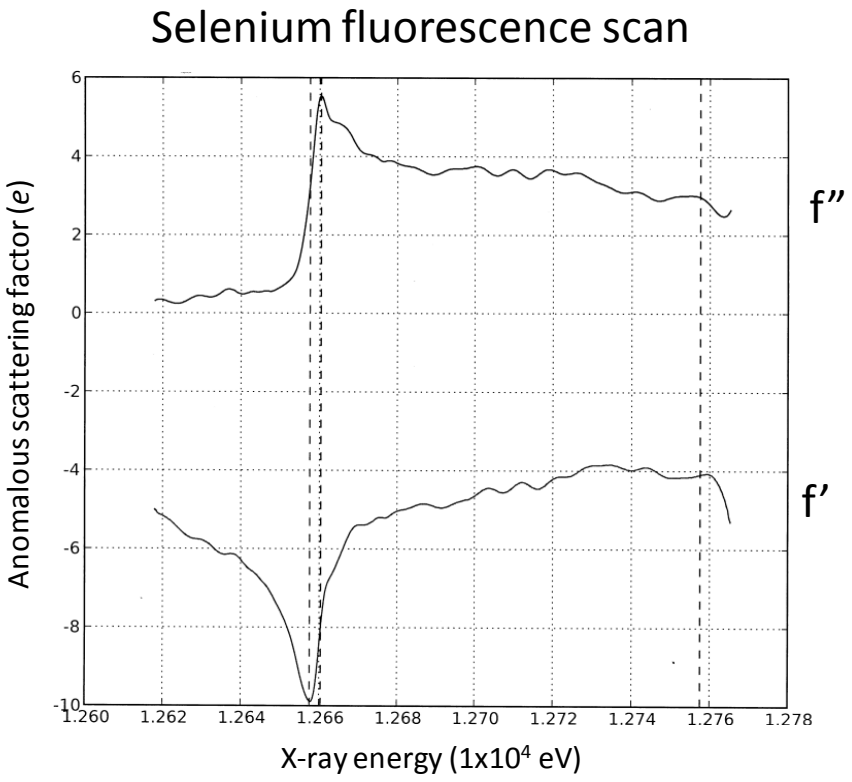


Figure 2.5 Fluorescence scan for selenium anomalous scattering

This X-ray energy scan of a selenomethionine-derivative crystal demonstrates the presence of a strong anomalous scattering signal at the selenium absorption edge. The most anomalous scattering is seen at 12.66 keV, which was the wavelength used for the SAD data collection.

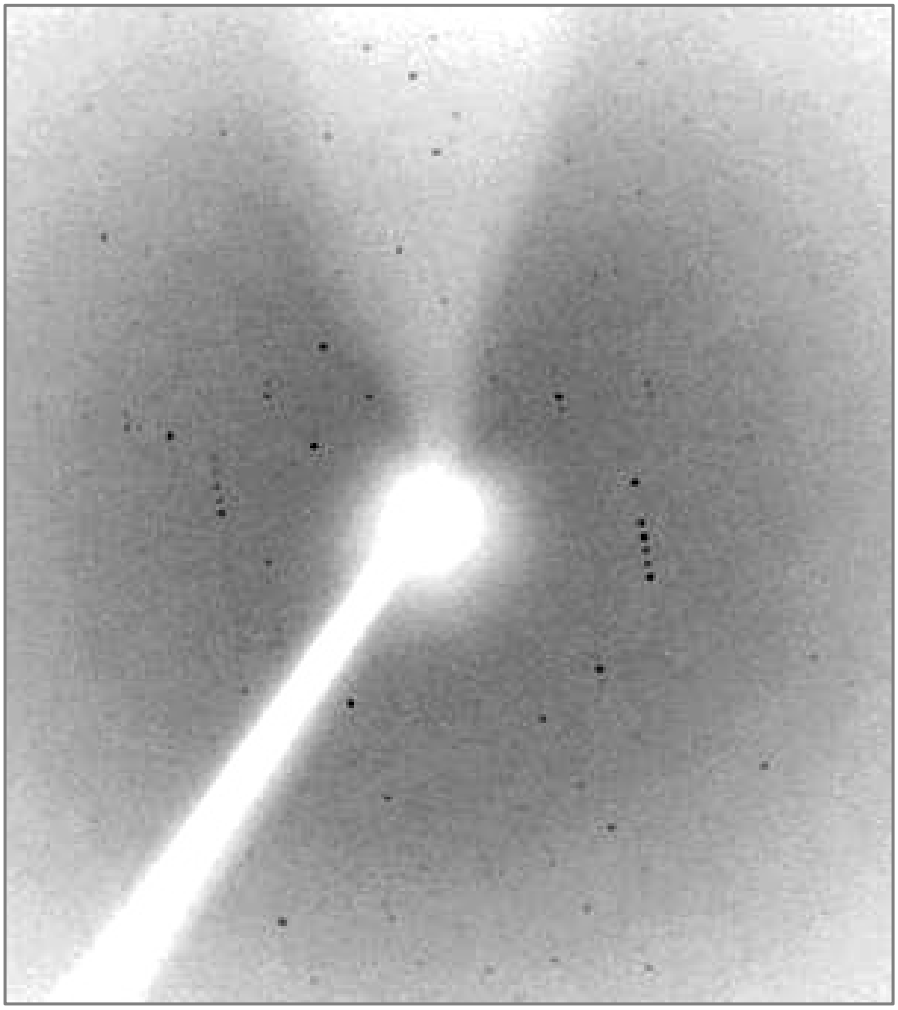


Figure 2.6 An X-ray diffraction image from a native Symplekin crystal

The X-ray diffraction pattern from this native crystal represents clear, well resolved, intense reflections that were used in data processing. Data was collected at the Advanced Photon Source, Argonne National Labs at the SER-CAT BM-22 beamline with a MarCCD detector. The reflections furthest from the center represent 2.2 Å resolution.

Table 2.2 Data collection statistics for X-ray diffraction of Symplekin 19-271

X-ray source	APS SER-CAT BM-22	
Space Group	$P4_12_12$	
Unit cell a,b,c (Å); α, β, γ (°)	68.7, 68.7, 138.5; 90, 90, 90	
Data set	SeMet	Native
Wavelength (Å)	0.97190	0.97958
Resolution (Å) (highest shell)	50.0-2.9 (3.0-2.9)	50.0-2.4 (2.49-2.40)
R_{sym}	9.4 (34.4)	8.0 (41.9)
I/σ	22.4 (1.0)	24.8 (1.9)
Completeness (%)	78.1 (6.7)	96.1 (79.6)
Redundancy	10.4 (1.6)	6.4 (2.8)

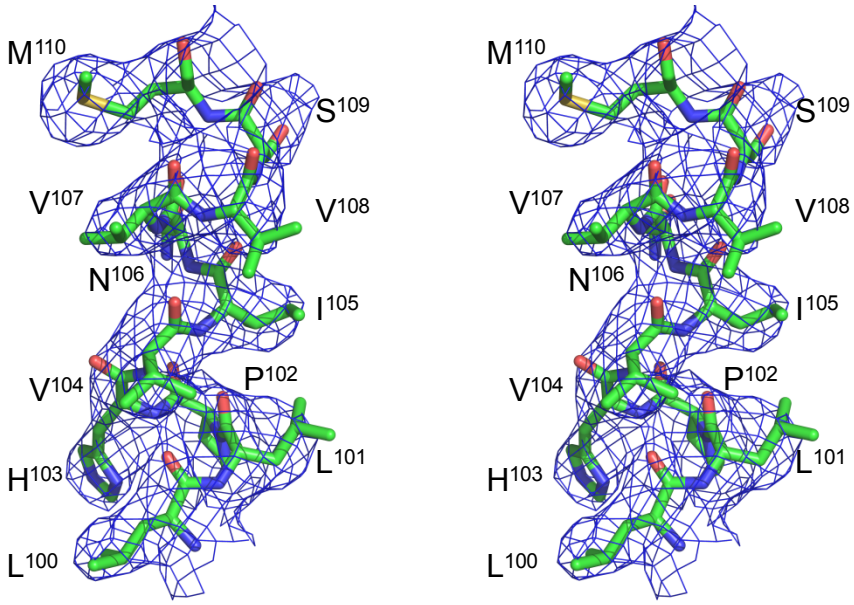


Figure 2.7 Wall-eyed stereo view of original experimental density with final model

Experimental density calculated from structure factors and phases by single-wavelength anomalous dispersion data covering the final model of Symplekin. The density is contoured to 1σ.

Table 2.3 Data processing and refinement statistics for Symplekin 19-271

Resolution (Å)	50.0-2.4
No. reflections	12465
R_{work}	0.2068
R_{free}	0.2653
Molecules per asymetric unit (AU)	1
No. of amino acids per AU	248
No. of waters per AU	142
Average *B*-factors	46.37
R.M.S.D. Bond lengths (Å)	0.0059
R.M.S.D. Bond angles (°)	1.20
Ramachandran (%) favored	96.76

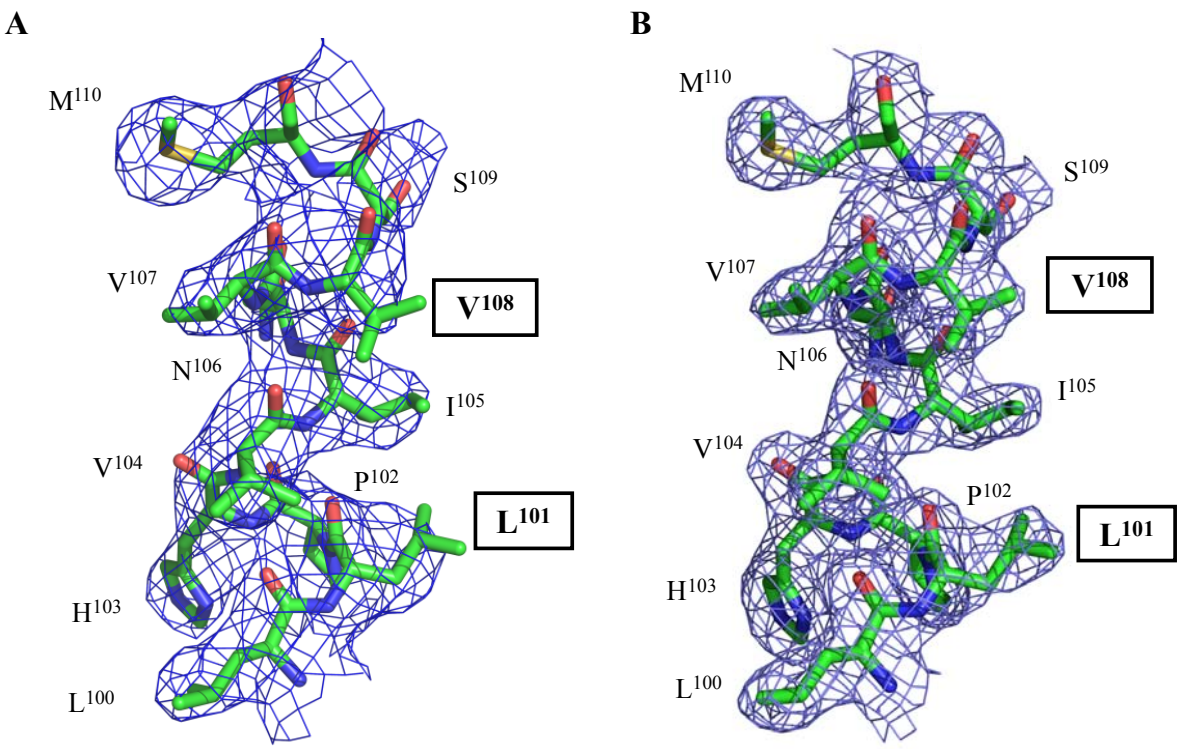


Figure 2.8 Density modified and refined electron density maps of a portion of the Symplekin structure

Simple electron density map contoured to 1σ around the final model of Symplekin residues leucine 100 through methionine 110. Letters corresponds to amino acid residues and numbers are consistent with the numbering of *D. melanogaster* Symplekin. Valine 108 and leucine 101 are boxed to show the dramatic effect of improved resolution. (**A**) Electron density map from 2.9 Å resolution selenomethionine derivative. (**B**) Electron density map from 2.4 Å resolution native data.

Chapter 3. Structural characterization of the Symplekin HEAT domain

3.1 Overall structural fold and interesting features

Residues 19-271 consists of a single domain with 5 bi-helical HEAT repeats forming an overall cradle-like shape with curvature of ~140° on the concave surface. This domain is named after four proteins that contain HEAT repeats: **H**untingtin, **E**longation factor 4, the PR65/**A** subunit of protein phosphatase 2A and the lipid kinase **T**OR. The 10 HEAT helices (residues 22-256) are lettered conventionally for HEAT repeat domains; helices composing the convex surface are labeled A, and concave surface helices are designated B (**Figure 3.1**). Each repeat is labeled in sequential number order with the short sixth helix labeled α6. Repeats 1-5 contain 37, 37, 47, 46, and 42 residues, respectively, falling into the range of residues established for HEAT repeats. Residues 257-270 form a short helix, perhaps leading into an additional unpredicted HEAT repeat, of which the first portion is observed in the crystal structure. The convex (A) helices are tilted slightly relative to the concave (B) helices. The interior of the protein forms a hydrophobic core, while the concave and convex surfaces are composed of mainly charged residues.

Symplekin residues 187-217 form an extended loop (loop 8) that caps the ends of α4B and α5B. A series of molecular salt bridges hold loop 8 in position (**Figure 3.2**). Bridges within the loop include the 2.8 Å bond between the backbone nitrogen of D192 and side-chain oxygen of S195, and the 2.9 Å bond between S203 backbone nitrogen and D206 side-chain oxygen. Residues involved in the six salt bridges between the loop and the HEAT

domain include helical bound residues M257 and R258 of α5B, K132 of α3B, and residues S195, G200, D201, and S203 of loop 8. All of these electrostatic interactions with loop 8 are shown in **Figure 3.2**. R258, D192, D201 and S203, which are involved in the salt bridges, are conserved among Symplekin homologues in the four higher eukaryotes and thus may play an important functional role. Comparisons of this loop with other HEAT domains indicate that it is a unique feature (see below).

3.2 Classification of individual HEAT repeats

HEAT domains are commonly found in solenoid structures that are protein scaffolds for multiprotein complexes[64-73]. Based on the specific amino acid positions in the structure, the HEAT repeats can be subclassified further into structural/functional families[64]. To sub-classify each HEAT repeat, the individual repeats were structurally aligned to HEAT repeat 2, by DaliLite[74](**Figure 3.3A**). Repeats 1, 2, 4 and 5 aligned well; however, repeat 3 is elongated and bent respective to the others, which is interesting because the unique loop 8 packs against this elongated helix as seen in **Figure 3.1**. Each of the repeats were compared to the sequence logo of three classes of HEAT sequences (ADB, AAA and IMB) reviewed in Andrade *et al.*[64]. All three classes, AAA, ADB, and IMB, contain conserved residues D19, V/I 24 and R/K 25. The AAA and IMB classes have a highly conserved P11 a hydrophobic residue at position 9. The ADB class lack P11, has a conserved hydrophobic residue at 28, and D/N/E at position 21. HEATs 1 and 5 are difficult to characterize because they are terminal repeats and have a different set of packing constraints[64]. HEAT repeat 2 contains the conserved residues common to all three classes: D77, V82 and R83. However, it lacks P11 and has N21 (N79), which indicates HEAT 2 belongs to the ADB class. HEAT repeat 3

includes D114, N115, I120, and K121, while HEAT repeat 4 contains 167D, 170N, 173I and R174, both lack prolines at position 11. Thus, all central repeats belongs to the ADB family. **Figure 3.3B** illustrates the conserved D19 and R25 present in both Symplekin HEAT repeats 2 and 4. This small ADB subclass contains α, β-adaptin and β-coat proteins that function as scaffolds for protein binding and transport. This sub-classification supports the hypothesis that the Symplekin HEAT domain has a structure appropriate for protein-protein interactions.

3.3 Amino acid conservation on the concave surface

Figure 3.4 shows the alignment of the N-terminus of Symplekin conserved residues among a wide range of species: *Homo sapiens, Xenopus laevis, D. melanogaster, Strongylocentrotus purpuratus, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Arabidopsis thaliana* (At1g27590). Previous published alignments only demonstrated the similarity in the center of the molecule[26], so this alignment demonstrates that the N-terminus of the Symplekin family is also conserved. Due to the diversity of these species, only five positions are 100% identical, but over the 247 positions in the HEAT domain, 96 are highly similar (defined as ≥6 species with similar residues). Within these 96 residues, 73 are nonpolar and are present to maintain the structural hydrophobic core of the HEAT domain. In general, the greater the species divergence the less the less the specific amino acid is conserved. Thus, when comparing the sequences of more closely related eukaryotes (*H. sapiens, X. laevis, D. melanogaster, and M. musculus*) (**Figure 3.5**), many more positions in the primary sequence are conserved.

Mapping conservation onto the three dimensional structure indicates regions that are the most likely to be important for function. The hydrophobic core of Symplekin is highly

conserved with 28 identical residues (**Figure 3.6**). By comparing the level of conservation in the concave, convex and loop regions of Symplekin, it is evident that the majority of identical residues fall on the concave surface and within loops (**Figure 3.7A**). The concave surface is lined with conservation, accounting for more than 20% of the total conserved residues in this HEAT domain. The convex surface contains only 4 conserved residues: R27, V31, M110, and E149. (**Figure 3.7B**). Examination of loop 8 in these four species shows that five residues are 100% conserved and seven positions contain highly similar residues (**Figure 3.5** and **Figure 3.7A, cyan**). Since residue conservation often implies functional importance, both the concave face and loop 8 are most likely important for Symplekin's biological function.

3.4 Symplekin HEAT domain structurally aligns with several other HEAT domains

Now that the conservation of the amino acids points to the functional importance of the concave surface and loop 8, other HEAT domains with closely related structures to Symplekin were investigated to understand how this domain typically interacts with binding partners. Other members of the HEAT domain family include serine/threonine-protein phosphatase 2A, Cullin-associated protein Cand1, and karyopherin α-1 subunit. Experimental evidence available for these proteins shows that their HEAT repeats are involved in protein-protein interactions and the majority of known HEAT repeat structures utilize their concave face as a binding or scaffolding surface[66,69,71,75-78]. To better understand the Symplekin HEAT domain and how it may interact with binding partners, DaliLite was utilized to align this domain with its closest structural neighbors: PP2A, karyopherin α, and Cand1[74].

Symplekin has the most structural similarity with human protein phosphatase 2A (PP2A) PR65/A subunit (1B3U, chain A); DaliLite aligned 104 Cα positions with a 1.7 Å RMSD and Z-score of 14.5 (**Figure 3.8**)[79]. PP2A PR65/A forms a horseshoe shape that provides a scaffold for the regulatory and catalytic domains of PP2A, a heterotrimeric enzyme critical to regulation of many cellular functions[71]. This protein has been classified as a HEAT domain and displays the classical cup shape for this type of repeat. The main areas of difference in the structural comparison are the elongated helices 4A and 3B and the extended loop 8 of Symplekin. Another interesting thing about this structural comparison is the much larger size of PP2A as compared with Symplekin. However, as seen in **Figure 2.1**, Symplekin is predicted to have much more alpha-helical character than what is seen in the N-terminal structure. Thus, by looking at proteins that encompass a HEAT domain, we can start to think about how the rest of Symplekin may be positioned relative to the N-terminus.

Symplekin structurally superimposes on the structure of yeast karyopherin-α with 11% identity over 196 Cα positions, a 5.0 Å RMSD, and a Z-score of 14.2. (**Figure 3.9**)[74]. The portion of karyopherin-α aligned with Symplekin has curvature of ~140° on the concave surface. The core of karyopherin-α is a canonical ARM repeat with an acidic concave surface equipped to bind the basic nuclear localization signal[77]. Symplekin aligns with the portion of karyopherin-α responsible for binding to nuclear localization signals (NLS), and both proteins share similar electrostatic potential on their concave face (**Figure 3.10**). *D. melanogaster* karyopherin-α3 binds to the NLS of HSF1, and it has been reported residues 1-124 of human Symplekin interacts with human HSF1[44,80]. The observation that karyopherin-α3 and Symplekin contain similar structural motifs, have similar electrostatic surfaces, and bind to HSF1 support the hypothesis that both Symplekin and karyopherin-α are scaffolds for

protein-protein interactions. However, with respect to loop 8, it is clear from the structural superposition of karyopherin-α and the Symplekin HEAT domain that the position of loop 8 clashes with the position karyopherin-α uses to contact NLS sequences. Thus, the details of complex formation may be distinct for the two proteins.

The third most structurally similar protein to Symplekin is Cand1 of the Cand1-cul1 complex. Cand1 is a protein responsible for scaffolding Cul1 and inhibits Cul1's ability to form the E3 ubiquitin ligase complex. Cand1 also uses its concave surface for protein binding (**Figure 3.11**). Again, loop 8 is unique to Symplekin. Examination of superpositions between the Symplekin HEAT domain and multiple other HEAT and ARM domain structures establish that, while many employ their concave surface for protein binding[74], loop 8 is unique to the Symplekin structure.

3.5 Structural implications for Symplekin in the mRNA processing complex

Residues 19-271 of *D. melanogaster* Symplekin form a canonical 5-HEAT repeat structure, with the exception of an extended loop 8 present in the Symplekins of known sequence. This is the first detailed structural information for any region of a Symplekin protein. Sub-classification of Symplekin's HEAT repeats and structural alignments indicate that Symplekin's structure is consistent with other proteins that act as protein scaffolds. Multiple interactions can occur on a HEAT domain platform; crystal structures and molecular dynamics studies of importin-β reveal four regions for peptide binding within 5 HEAT repeats [81]. Examination of the electrostatic potential of the HEAT domain shows that the concave surface is positively charged; indicating this surface is poised to bind negatively charged proteins. In contrast, the ridge formed by even-numbered loops, including loop 8, is

negatively charged, thus allowing interaction with positively charged proteins (**Figure 3.10**). This is further evidence that this Symplekin HEAT domain is designed to accommodate multiple, diverse partners.

As mentioned in the Introduction, the N-terminal HEAT domain of Pta1 (yeast Symplekin) binds both Ssu72 and Glc7p[33,45] and human Symplekin uses its N-terminus to bind HSF1[44]. Disruption of these interactions results in the abrogation of 3'-processing. Depletion of Glc7p causes an accumulation of phosphorylated Pta1 and reduction of polyadenylation, but can be rescued by adding the Glc7p phosphatase back into the processing reaction[33]. Symplekin binding to Ssu72 promotes proper 3'-end processing and binding to HSF1 promotes polyadenylation of Hsp70 mRNA in heat stressed cells[44,45]. Much of the literature regarding Symplekin suggest that it is a scaffold for assembling functional 3'-processing machinery, namely CPSF and CstF subunits, but the interactions directly involving the HEAT domain appear to have more of a regulatory purpose. Thus, we propose that the N-terminal HEAT domain of Symplekin acts as a regulatory scaffold for 3' processing through interactions with phosphatases Glc7p and Ssu72 or transcription factors such as HSF1, while the C-terminal region is available for scaffolding and positioning the 3'-end processing core proteins.

Structures of nearly all of the other core components of the 3'-end processing machinery have been determined, including CPSFs 30, 73, 100, CstF 64 and 77 and CF I_m-25[16,82-84]. The electron microscopy image of purified 3'-end processing complex including Symplekin, CPSF, CstF and CFI was recently determined[28]. This 50S complex displays a maximal dimension of 25 nm and exhibits a kidney shape, elongated and bent with two areas of high density and a cavity between them. This length is likely dependent upon the

structural scaffolds within the processing complex, which literature shows are Symplekin and CstF 77[2,85]. The mouse CstF 77 structure exhibits a TPR domain, a domain similar to Armadillo and HEAT domains[82]. CstF 77 uses its C-terminal Pro-rich region to bind to the CstF 50 and CstF 64, thus leaving the long TPR region available for binding other processing proteins[26]. The CstF 77 TPR dimer is 16.5 nm in length[2]. Symplekin HEAT domain has a length of 6.5 nm, 25% of the total length of the EM image mRNA processing complex. The armadillo C-terminal portion of Symplekin is likely much longer because it nearly twice the size of the HEAT domain (**Figure 2.1**). The length and known binding partners of these two proteins support the idea that they make up the 25 nm length of the mRNA processing machinery visualized by EM. Detailed crystal structures, electron microscopy images and small angle x-ray scattering of the 3'-end processing complex are all necessary for obtaining a complete understanding of the structure/function relationship within this macromolecular complex; the Symplekin HEAT domain reveals one more detailed piece of this structural puzzle.

3.6 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 3.

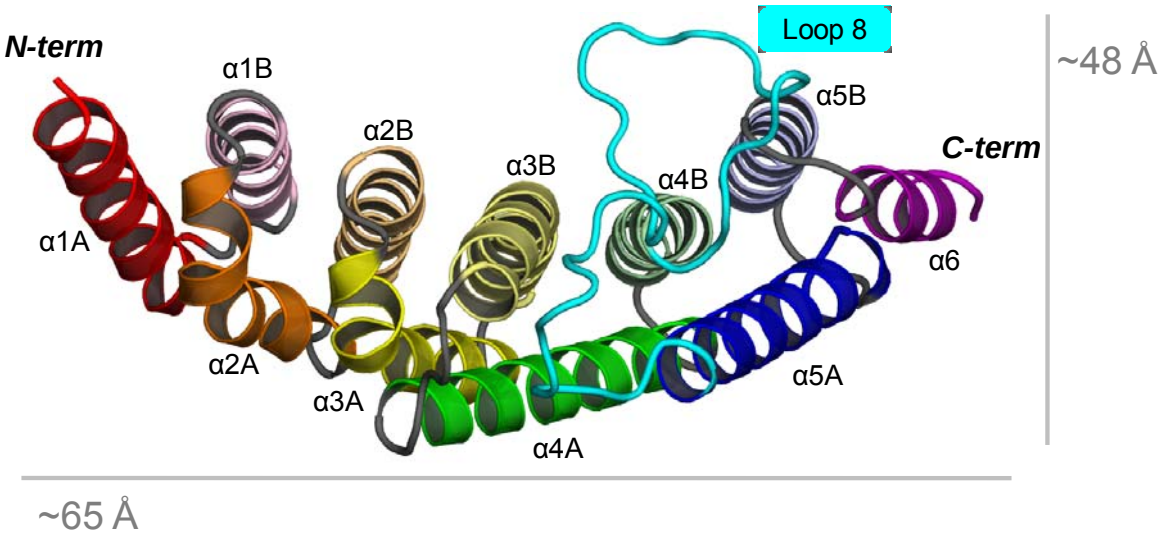


Figure 3.1 Overall structure of the Symplekin HEAT domain

Illustration of the HEAT domain of the N-terminal region of Symplekin. The termini are labeled and the coloring is in rainbow (red, orange, yellow, green, blue, and purple) from N to C terminus. HEAT repeat helices (A and B) are in the same color, but with the concave helix in a lighter shade. Helices are assigned letters and numbers based on the accepted method for naming HEAT repeats. A helices make up the convex surface and corresponding B helices in each repeat make up the concave surface. Loop 8 is shown in cyan, while all other loops connecting intra- and inter-HEAT repeats are gray.

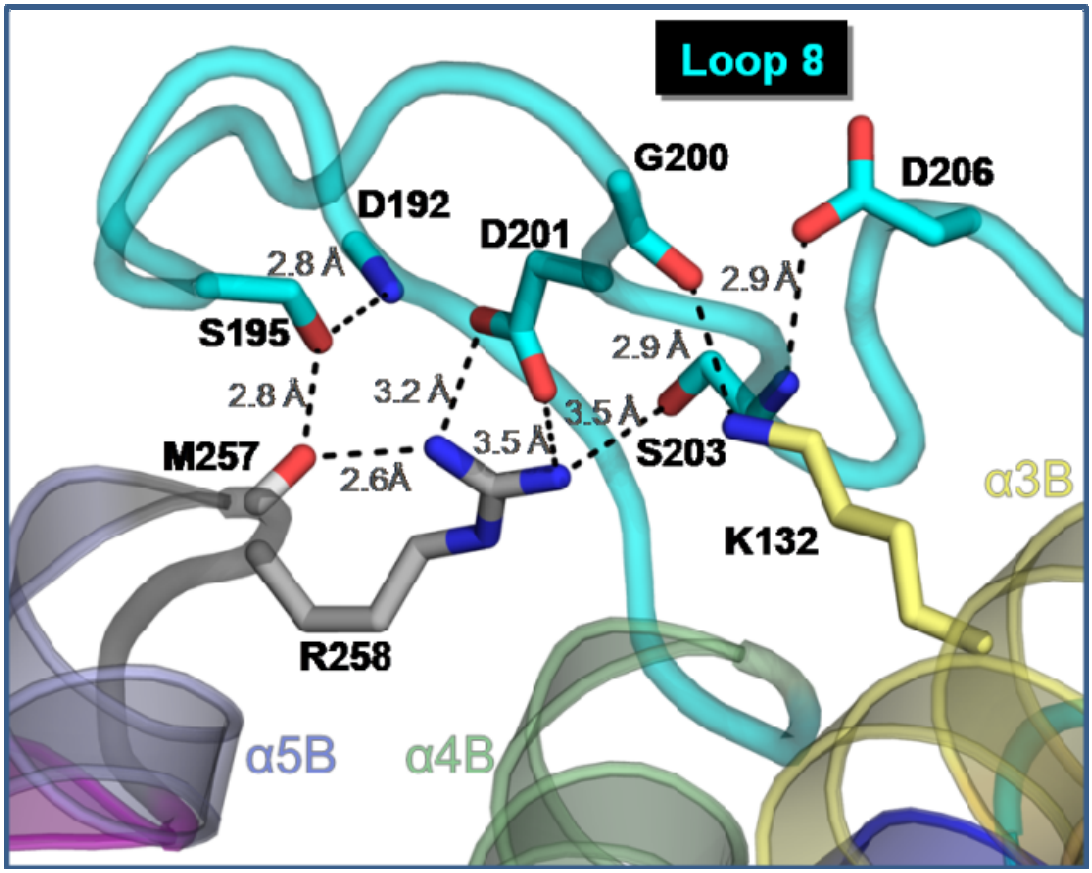


Figure 3.2 Extensive electrostatic interactions position Loop 8 over concave surface

Loop 8 is held in position by many electrostatic interactions. Concave surface residues R258 and K132 use their side chain nitrogen atoms to form contacts within Loop 8 through D201/ S203 and G200, respectively. M257 also contacts Loop 8 through S195. Many contacts within the loop also appear to be important to maintain its fold.

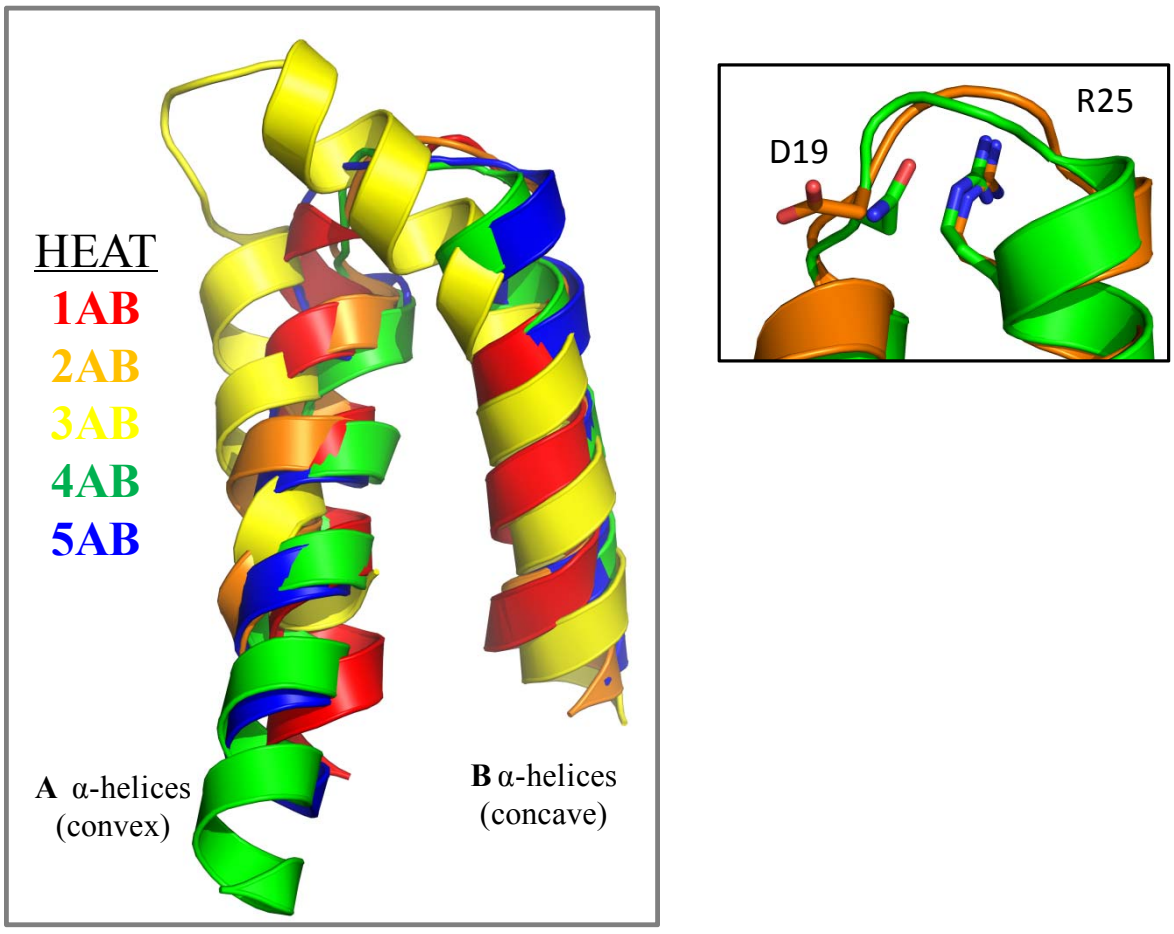


Figure 3.3 Alignment of individual HEAT repeats within the Symplekin HEAT domain

(**A**) Structural alignment of each HEAT repeat within Symplekin. Helix 3B is much longer and has a unique bend relative to the other B-helices. (**B**) D19 and R25 of repeats 2 and 4 are residues well conserved in HEAT repeats and help define the structure as a HEAT domain.

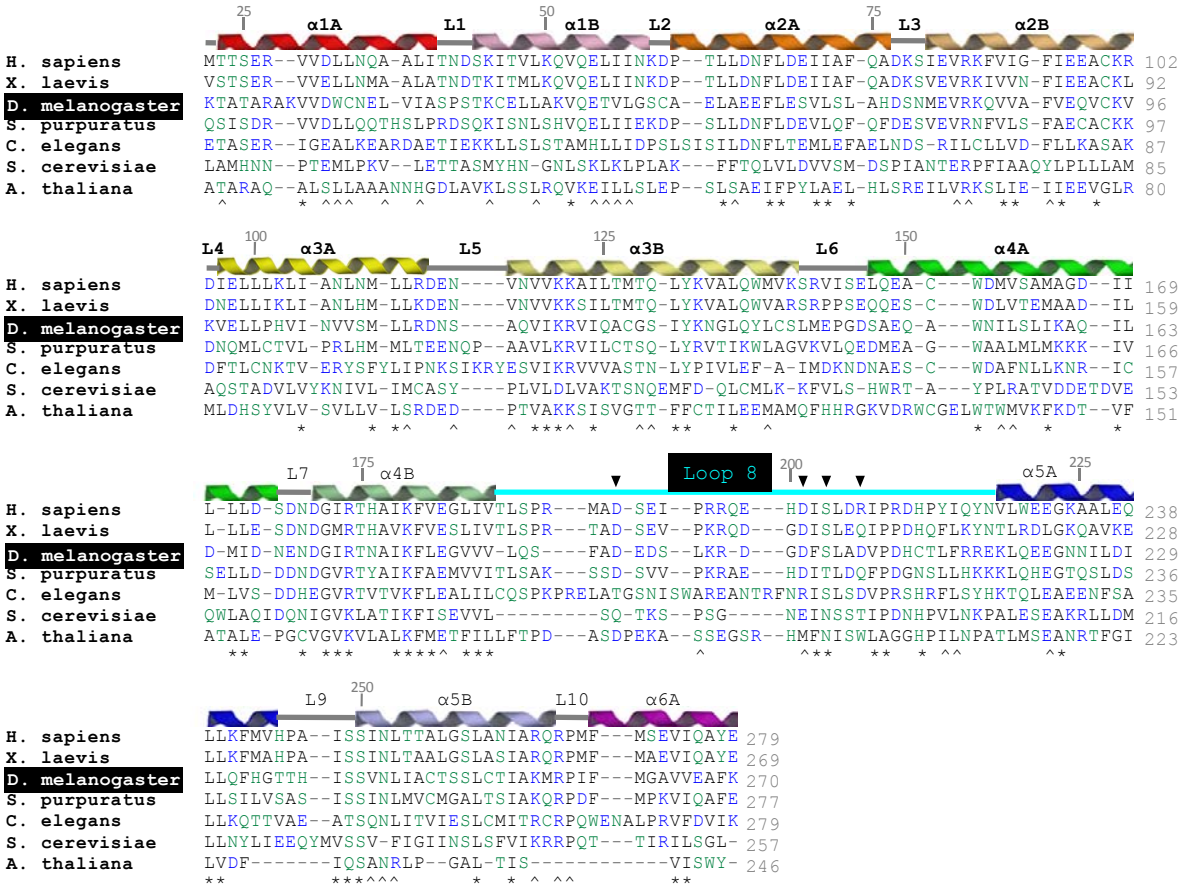


Figure 3.4 Sequence alignments of Symplekin homologues from diverse species

Sequence alignment of the N-terminus of several Symplekin homologues annotated with the secondary structural elements resolved in the structure of *D. melanogaster* Symplekin HEAT domain. Species include *H. sapiens*, *X. laevis* (frog), *S. purpuratus* (sea urchin), *C. elegans* (worm), *S. cerevisiae* (yeast), and *A. thaliana* (plant). Polar residues are in green, charged polar residues are blue, hydrophobic residues are black. Black triangles denote residues found to make electrostatic interactions with concave surface residues in the crystal structure. Numbers on the top of the sequence correspond to the *D. melanogaster* residues.

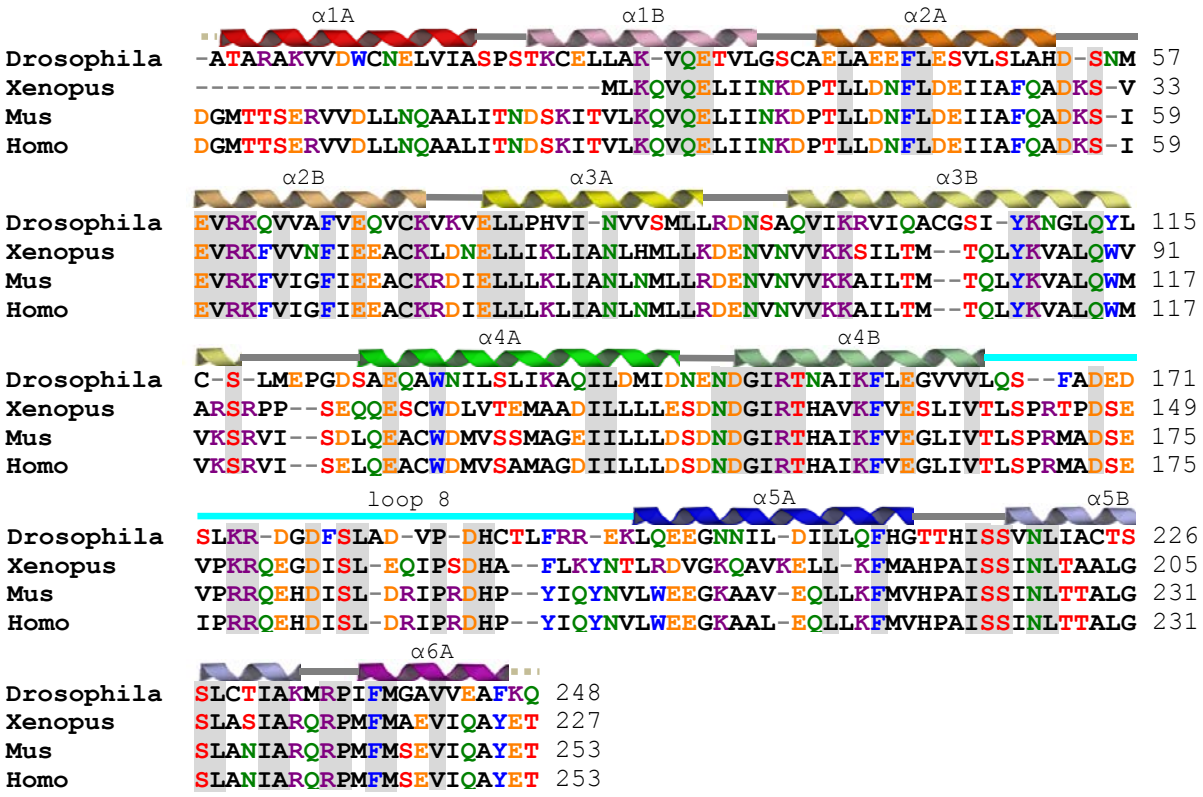


Figure 3.5 Sequence alignment of *D. melanogaster* Symplekin with higher eukaryotes

Sequence alignment of *D. melanogaster* Symplekin protein with *X. laevis*, *M. musculus and H. sapiens.* Secondary structural elements corresponding to the HEAT domain of N-terminal Symplekin are listed above corresponding residues. Gray bars indicate areas of conservation.

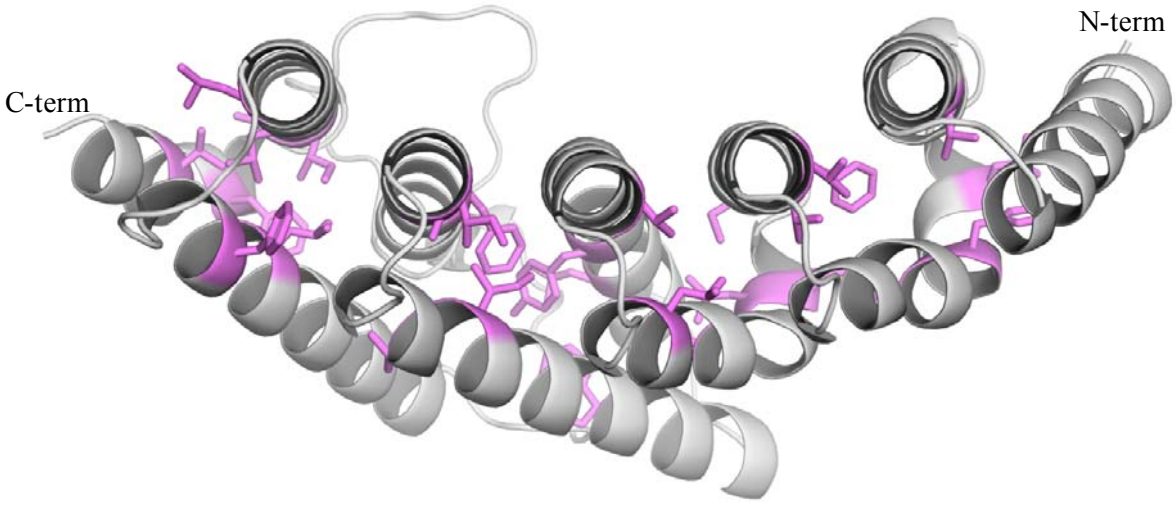


Figure 3.6 Conserved residues in the hydrophobic core of Symplekin

Residues shown in stick representation and colored purple are residues that are highly conserved among Symplekin homologues (See Figure 3.5) and that fall into the hydrophobic core of the N-terminal Symplekin HEAT domain. All other regions of the protein are gray and the termini are labeled. The hydrophobic core is essential for maintaining the bi-helical fold of HEAT repeats.

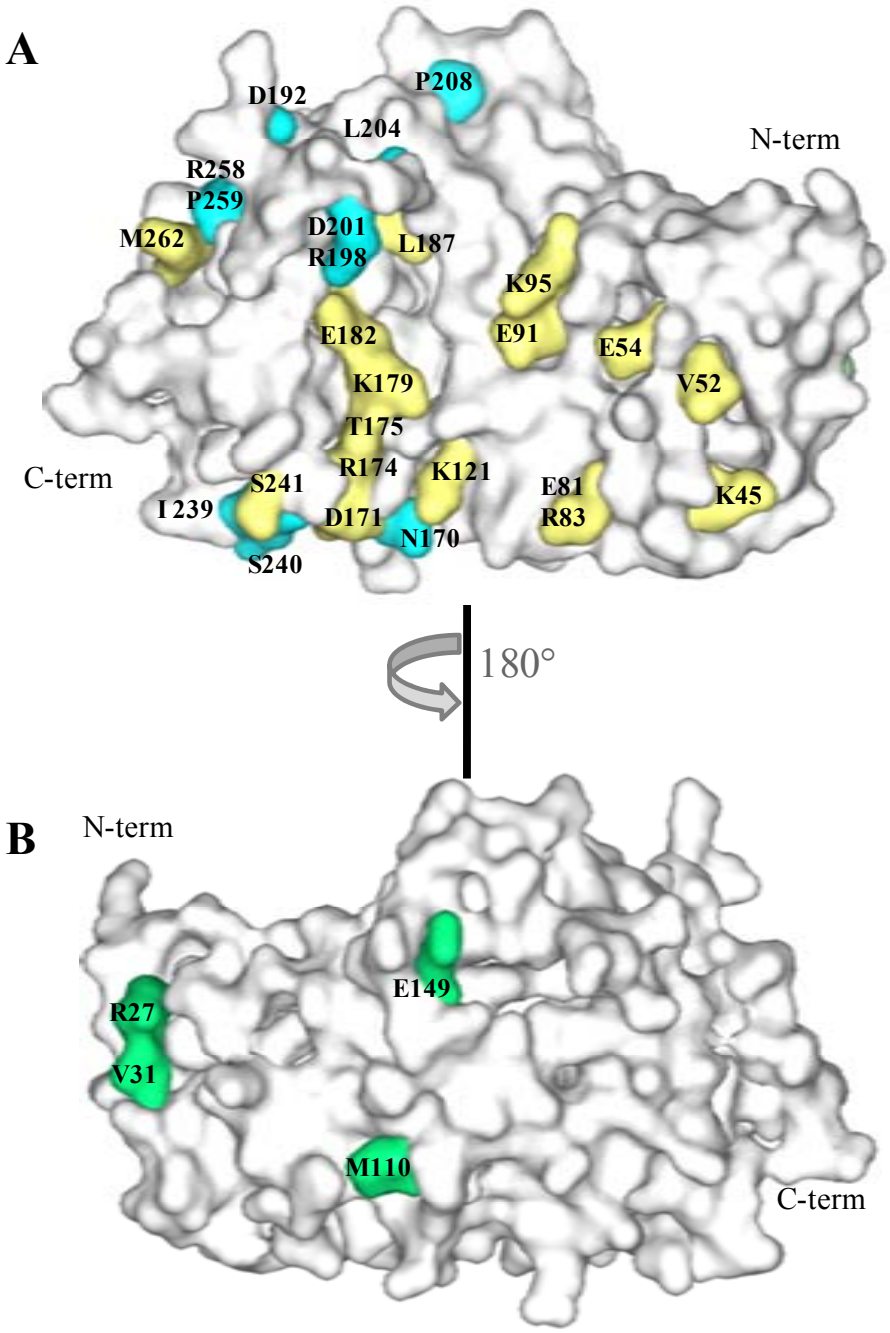


Figure 3.7 Conserved residues on the concave and convex surfaces of Symplekin

The surfaces of Symplekin are rendered with color representing conservation. (**A**) Looking at conserved residues projecting from the concave surface, yellow on helices, cyan in loop regions (**B**) Looking at the convex surface, green residues are conserved. The concave surface contains many more conserved residues than does the convex surface. Also, loop 8 contains many conserved residues. This suggests that the concave surface may be more important than the convex surface for Symplekin's biological scaffold function.

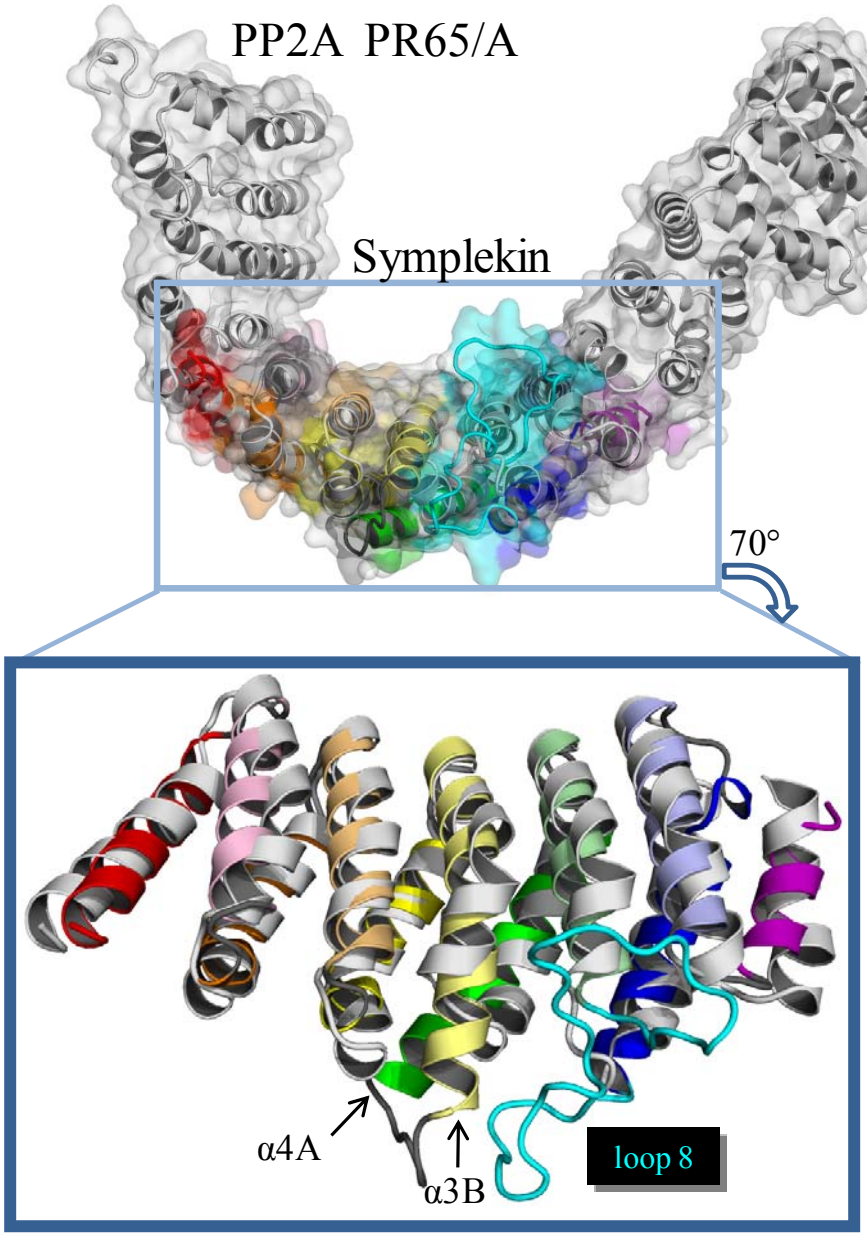


Figure 3.8 Symplekin structurally aligned with PP2A regulatory domain

Symplekin (colored as in Fig. 3.1) overlaid with PP2A PR65/A regulatory domain in gray (PDB 1b3u, chain A). A closer inspection of the area where Symplekin aligns shows that loop 8 is unique from the other loops in both structures. Also, α4A and α3B have extended helices relative to the other HEAT repeats. Loop 8 packs against this longer α3B helix. PP2A regulatory domain is a scaffold for the PP2A phosphatase activity, and it's structural similarity to Symplekin may indicate that Symplekin can also provide this regulatory scaffold function.

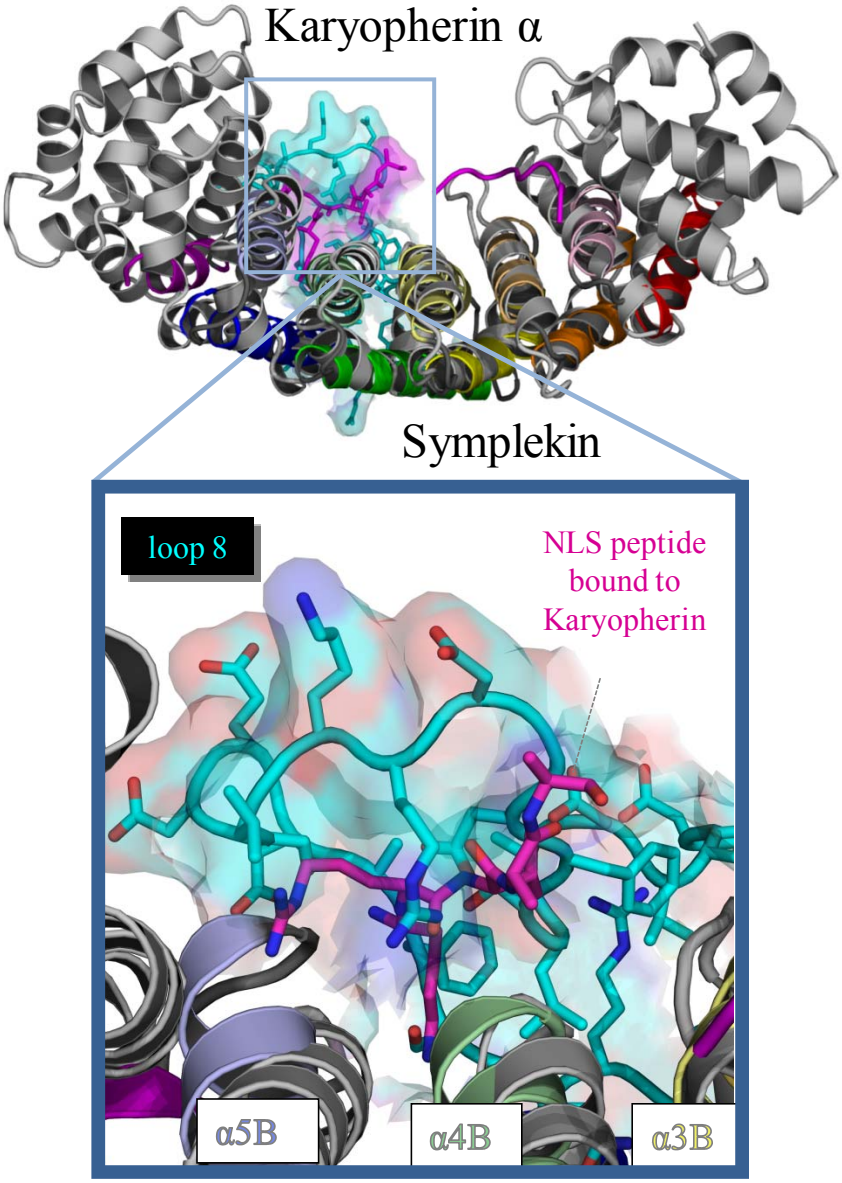


Figure 3.9 Symplekin structurally aligned with Karyopherin-α

Symplekin shown in rainbow, Karyopherin-α in gray, nuclear localization peptide bound to Karyopherin-α surface in magenta. Close inspection of this structural alignment indicates that Symplekin loop 8 (cyan) lies in the same region that is Karyopherin uses to bind to nuclear localization peptides. This overlay illustrates the ability of the concave surface of these types of structural domains to be used as binding surfaces.

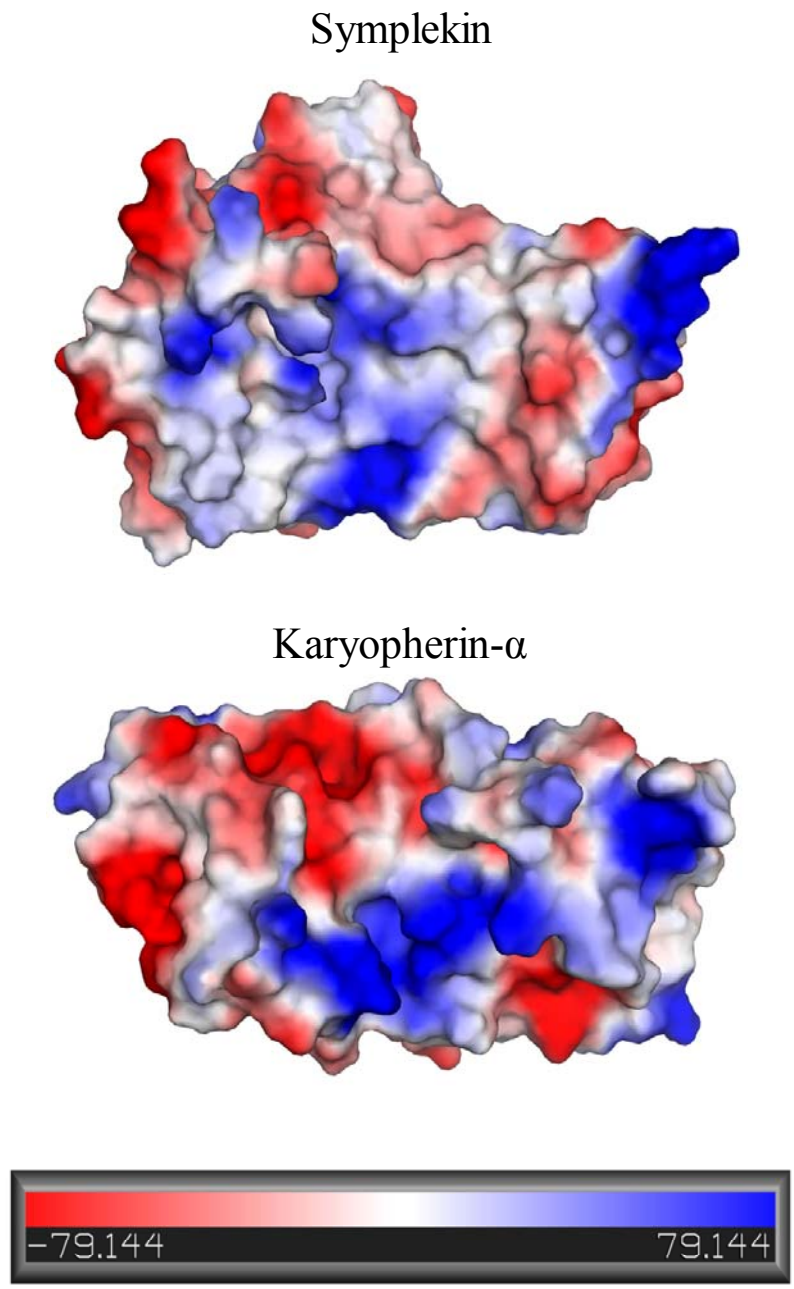


Figure 3.10 Electrostatic potential of the concave face of Symplekin and karyopherin-α

Red areas indicate regions of negative charge, while blue regions correspond to areas of positive charge. The ridge formed by loop 8 and the other even numbered loops in Symplekin have a negative charge (red), while the concave surface is mainly positively charged (blue). The region of negative charge (red) on karyopherin-α is used to bind positively charged, lysine/arginine-rich nuclear localization signals. The distribution of charge on Symplekin's surfaces may be important for its ability to bind to multiple diverse proteins.

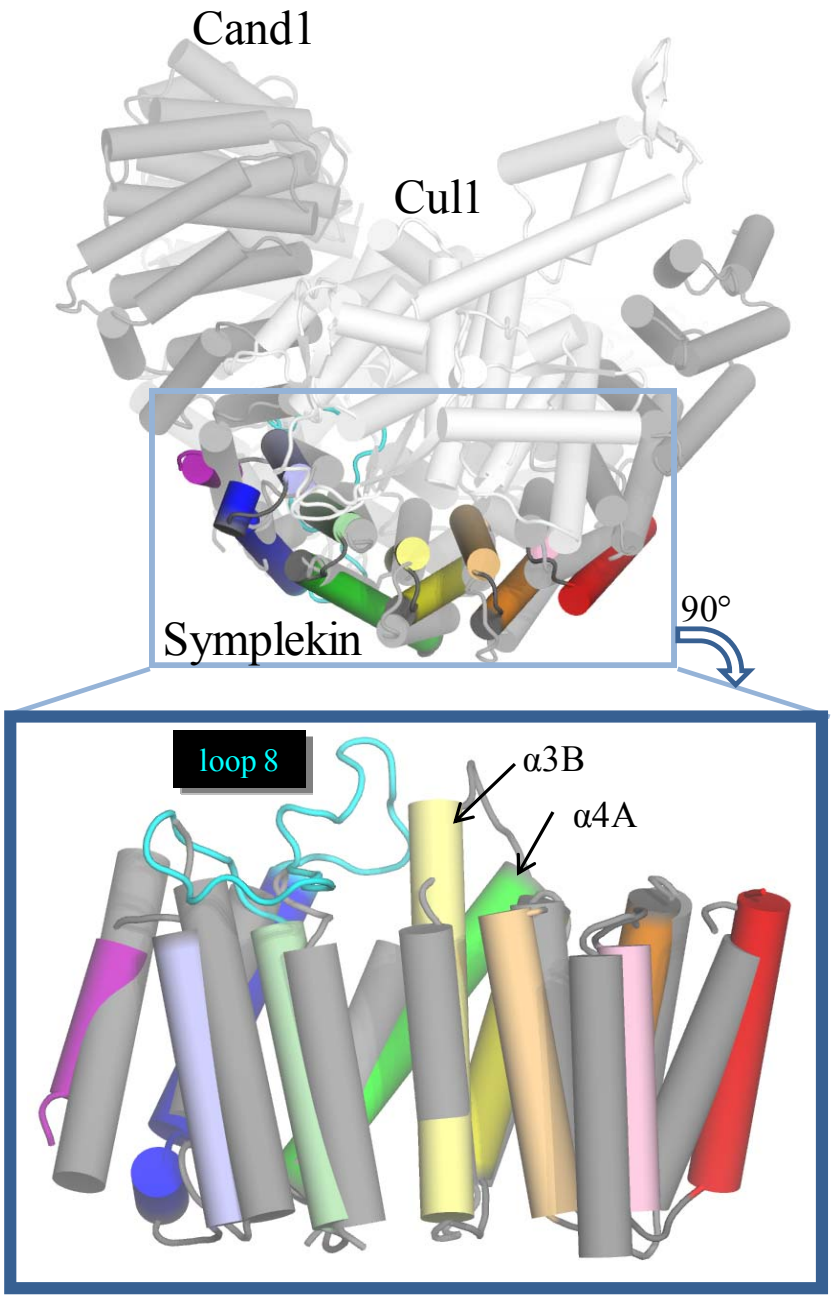


Figure 3.11 Symplekin structurally aligned with Cand1

Helical representation of Symplekin colored as in Fig 3.1, Cand1 in gray, Cul1 in white. Rotating the structure 90° forward and cutting away the regions that don't align; the structures show that loop 8 is a unique extended loop that lines up against helix α3B. As in Figure 3.8 and 3.9, it is clear that HEAT domains can be used to scaffold other proteins and the extended loop 8 is unique to Symplekin.

Chapter 4. Biophysical characterization of the Symplekin HEAT domain through molecular dynamics simulations

4.1 Why use molecular dynamics to study Symplekin?

Molecular dynamics (MD) simulations have become an increasingly popular method to understand how a molecule's motion is tied to its biological function. Specifically, simulations with exportin Cse1p, a HEAT domain protein, provide a picture of a possible intermediate state between the crystal structures of the RanGTP bound and unbound states[72]. Another MD study shows that importin-β unfolds from it's 9 nm length to 15 nm length when its ligand is removed[73]. Focusing on the hydrophobic core of armadillo and HEAT repeats, Parmeggiani *et al*. utilized molecular dynamics to find the most energetically favorable core and modular peptide-binding motifs[86]. Each of these examples provides evidence for the valuable use of molecular dynamics simulations to gain an understanding of how energy and motion are intimately tied to the biological function of macromolecules.

Through solving the structure of the Symplekin HEAT domain, I have demonstrated that this protein indeed has the structure of a molecular scaffold. This scaffolding role of Symplekin was predicted because Symplekin can bind to a variety of proteins and has been implicated in regulating the orientation of the CPSF and CstF complex for 3'-end mRNA processing (see **Figures 1.1, 1.3** and **Table 1.1**). I sought to strengthen the argument that Symplekin is a molecular scaffold by studying its motion through MD. I used MD to investigate the overall magnitude of motion for the HEAT domain and also to appreciate how motions within this domain are correlated to each other. Second, I wanted to probe the

importance of the conserved loop 8 region of the molecule and observe its effect on the overall motion of the molecule. These MD studies help to explain how the HEAT domain of Symplekin is energetically poised to bind to multiple protein partners.

Molecular dynamics is a method that uses computer simulations that are guided by laws of physics, mathematics and chemistry. An initial velocity and position is assigned to each atom in the molecular model, and the interaction (force and position) between all atoms is monitored over a given time period. Nanosecond time scale simulations allow us to see how a molecule breathes. Femto- through pico-second timescales are used to see bond stretching and angle bending, while pico- through nanosecond timescales are the timescales used to see surface side chain motion, loop motion and collective motion. The theory of MD is based on Newtonian physics and statistical mechanics. I am not an expert in either physics or statistical mechanics, so I collaborated with Monica Frazier on the MD studies of Symplekin. She has studied these topics and has published a peer-reviewed article with Matt Redinbo using this technique.

4.2 Design of Symplekin mutants for molecular dynamics

Three initial structural models were utilized in the molecular dynamics simulations. These include wild-type, short modeled loop 8 and poly-ser loop 8, as illustrated in **Figure 4.1.** Only the region of loop 8 was modified in these structures, the rest of the molecule remained the same. The reason for making the two mutant forms of loop 8 was to probe the importance of the presence of the loop, and also the presence of specific charged residues within the loop. These simulations will demonstrate the importance of this conserved feature on the energetic profile of the molecule.

To understand how the presence of loop 8 affects the magnitude and correlation of motions in the Symplekin HEAT domain, loop 8 was replaced with a loop having characteristics of other inter-helical loops in HEAT domains. To design a linker that would not disrupt the overall motion of the core HEAT helices, the distance between the Cα termini of α4B and α5A was determined and maintained in the mutation; the length between leucine 187 and arginine 216 of wild type Symplekin is 10.6 Å. To select a linker, many other HEAT and Armadillo proteins were examined to identify common inter-helical linkers of 10.6 Å length. Six residues were repeatedly used to bridge a 10.6 Å gap in HEAT repeats. To keep this linker as authentic as possible, native residues were reserved on each end of the loop and connected with a residue common to most loops, glycine. The linker was designed and mutated *in silio* using COOT[61] as 187-LQSGRR-216. This short linker connects α4B with α5A and replaces the lengthy charged loop 8. During the following discussion, this is referred to as the "short loop 8" and is considered to be a conventional loop to connect HEAT repeats.

The second mutation of loop 8 was designed to study the importance of specific ionic interactions both within the loop and between the loop and concave surface. Thus, polarity was maintained, but the specific ionic interactions were disrupted by mutation of charged residues to serine. D192, E193, D194, K197, R198, D199, D201, D209, H210, R215 were all computationally mutated to serine. This mutant will be referred to as the "Poly-Ser loop 8". Each model was subject equilibration before MD; this step will allow the models to find a low energy equilibrium state[87].

4.3 Molecular dynamics simulations methods

In collaboration with Redinbo lab member Monica Frazier, molecular dynamics simulations of the Symplekin HEAT domain, including wild type, modeled loop 8 and poly-ser loop 8 were performed using the AMBER 2003 force field with at 2 fs time step, with a total run time of 15 ns[87]. Several programs within AMBER were utilized for setting up the simulations. LEaP was used to generate the topology and parameter files, SANDER performed the 5000 steps of energy minimization, which included constant volume followed by constant temperature equilibration, the PMEMD module was used for the production runs, and PTRAJ was utilized for analysis of the results[87]. TIP3P water molecules were used to generate the solvated structure[88], and electrostatic interactions were calculated using the particle-mesh Ewald algorithm with a cutoff of 10 Å applied to Lennard-Jones interactions[89]. All molecular dynamics simulations were modeled after the previous simulations performed in the Redinbo laboratory[90].

4.4 Result of molecular dynamic simulations support Symplekin's role as a scaffold

The first step in analysis of the MD simulations is to make sure that each simulation reached equilibrium. Root mean squared deviations (RMSD) of C-α peptide backbone atoms (**Figure 4.2A**) and all atoms (**Figure 4.2B**) do not fluctuate significantly after 5 ns in any of the simulations. Thus, the data between 5-15 ns were utilized for analysis, and the first 5 ns were discarded. Examination of the atomic position fluctuation shows that the loop regions of the protein have the largest change in position, especially loop 8 (residues 187-216) (**Figure 4.3**). It is not surprising that most of the residues in the helices have lower

fluctuations because they are held in place by an extensive hydrophobic network in the interior of the protein, as can be seen in **Figure 3.6**. The overall magnitude of Cα atomic position fluctuations (apf) $Å^2$ for each simulation confirms that the wild-type and poly-ser models have a similar level of motion, but the short modeled loop has a 20% increase in average fluctuation (**Table 4.1**). The apf was calculated with and without the terminal residues because untethered termini typically exhibit more fluctuation. In this case, however, the termini did not significantly add to the overall motion. Upon closer inspection, the residues in the canonical modeled short loop have a 0.5 $Å^2$ increase in apf compared to the poly-ser or wild-type loops, even thought the short loop is only composed of 6 residues versus the 29 residues in the wild-type loop. The presence of loop 8 is dampening the level of motion throughout the whole molecule.

Besides looking at the magnitude of motions between the simulations, we can also examine the way the movements are correlated within the protein by examining correlation plots. Correlation plots have each residue plotted on the x and y axis, so that the interdependence of movement can be compared between each set of residues. The plot of wild-type Symplekin residues shows correlated movements in the core of the protein and anti-correlated motion at the termini (**Figure 4.4**). In other words, the core of the protein moves together as one unit, while the termini are able to move in opposite directions from the core. The modeled short loop Symplekin simulation shows the same areas of correlated and anti-correlated motion, but the level of correlation has greatly increased (**Figure 4.5**). Increased correlated movement means fewer degrees of freedom for the protein to sample conformational states. Since Symplekin has multiple binding partners, the motion within the molecule should allow docking of many different proteins. Thus, it appears that loop 8 is

dampening the overall correlated movements in this HEAT domain, which allows this scaffold flexibility for its diverse binding partners. The wild type and poly-ser Symplekin correlation plots display the same areas and levels of motion (**Figure 4.6** and **Table 4.1**). This indicates that the presence of lengthy loop 8 dampens the overall motion of the HEAT domain, but specific residues in the loop are not essential for this dampening motion.

4.5 Key electrostatic interactions are maintained during MDS

To further understand the role of loop 8, the electrostatic interactions holding the loop in place in the crystal structure (**Figure 3.6**) were examined throughout the dynamics simulations. Over each 0.002 fs time step in the simulation, the distance between each set of atoms making electrostatic contacts in the crystal structure was examined. **Table 4.2** lists each of the atomic distances that were investigated in the wild-type simulation. The distances monitored over the simulation time-course are shown in **Figures 4.7, 4.8 and 4.9.** (Time shown in figures represents the 10 ns after the 5 ns equilibration time. See beginning of this section for explanation.)

K132, the only residue on alpha helix 3B that has electrostatic interactions with loop 8 in the crystal structure, does not maintain these contacts during the simulation (**Figure 4.7A**). However, the G200-K132 residues stay within 5 Å for the last 8 ns of the simulation, after their distance reaches a maximum of 11 Å. This implies that this part of the loop maintains close proximity with helix 3B. Loop 8 S195 does not maintain close contact with helix 5B M257, nor does loop 8 S203 and helix 5B R258 (**Figure 4.7 B, C**). However, since the overall apf for loop 8 is low, there is not a large movement away from the position seen in the crystal structure.

Select interactions between residues within loop 8 do maintain close proximity during the simulation (**Figure 4.8**). These include serine 203 interacting with aspartic acid 206 and aspartic acid 192 interacting with serine 195. These observations indicate that loop 8 is moving in a directed manner and some interactions within the loop may be important in maintaining its position for its protein binding functions. This could be tested by making these four residues nonpolar and performing protein binding assays. Also, residues arginine 258 and methionine 257 in loop 10 maintain proximity through an electrostatic interaction, indicating that the backbone oxygen on methionine 257 is purposefully projecting arginine amine groups towards loop 8. In this way, the core HEAT domain is communicating with loop 8.

Another interesting observation about arginine 258 is that it maintains close proximity to aspartic acid 201; this contact seen in the crystal structure is maintained throughout the simulation (**Figure 4.9**). Both aspartic acid 201 and arginine 258 are conserved among Symplekin homologues in divergent species (**Figure 3.4**). Mutations of conserved residues D201 and R258 would be very interesting to monitor in a protein binding assay. As discussed in the following chapter, Dr. Mindy Steiniger is continuing to develop a biochemical assay for testing protein-protein interactions with Symplekin. Surely, abolition of this key electrostatic interaction would have an impact on the loop's contribution for protein binding. (Replications of each of these molecular dynamics simulations is currently underway and will be incorporated into the final version of this dissertation.)

4.6 Molecular dynamics summary

Loop 8 dampens the overall motion and the correlated motions within the Symplekin HEAT domain. These results suggest that loop 8 encourages a more stable binding surface for protein interactions. Also, loop 8 increases the degrees of freedom for movement correlation, improving the HEAT domain's flexibility to bind to multiple partners. Key interactions between loop 8 residues and the concave helices are maintained and no large movements of the loop are seen during the simulations, indicating that it may maintain it's proximity to the concave surface to for its biological function as a protein scaffold.

4.6 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 4.

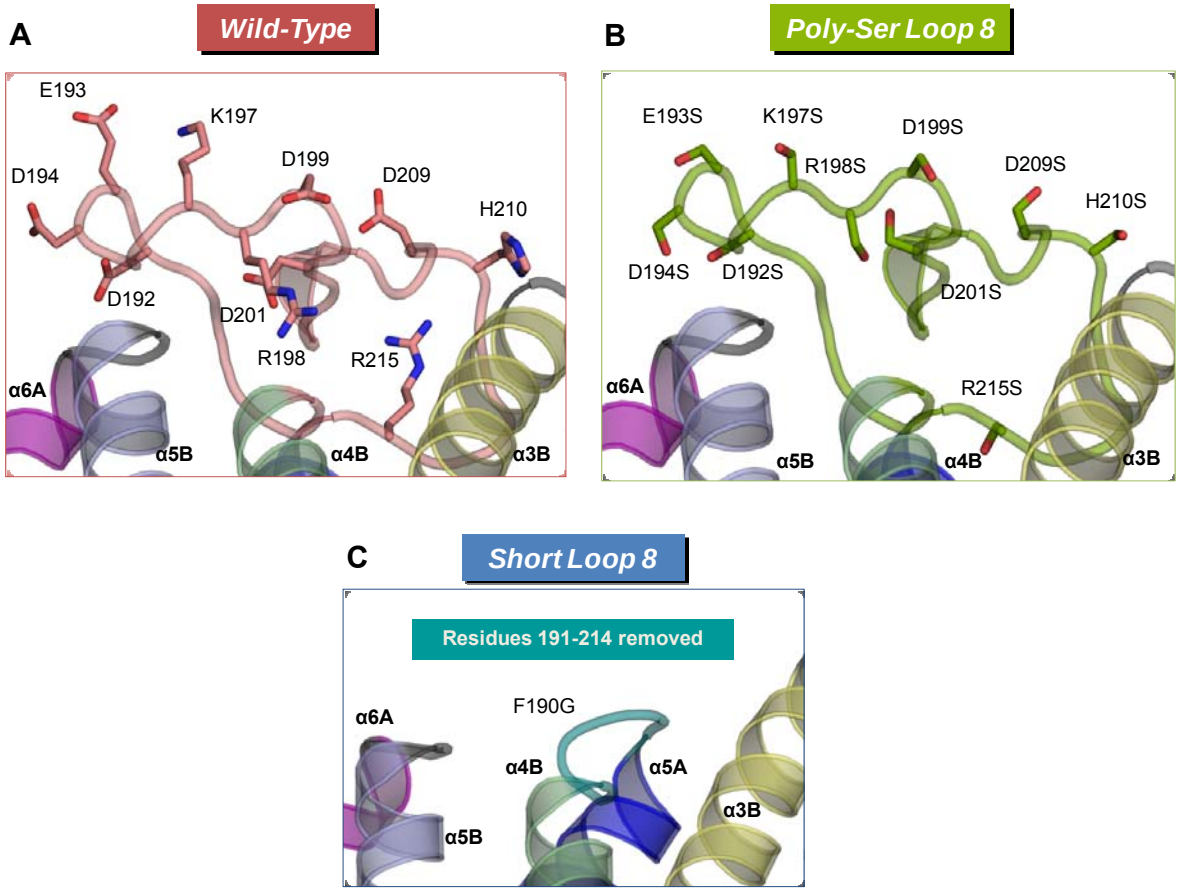


Figure 4.1 Loop 8 region variants of the Symplekin HEAT domain used in three independent molecular dynamics simulations

(A) Wild-type Symplekin with native residues in loop 8 region. The residues that are later mutated to serine are shown in stick representation. Helices are labeled and colored the same way as previously described in Figure 3.1. (**B**) Poly-Ser loop 8 modeled by changing loop 8 residues shown in sticks into serine. (**C**) Short loop 8 model prepared by removing residues 191-214 and joining native residue 189 and 215 by placing a glycine residue at position 190. This represents a canonical inter-helical loop found in HEAT domains.

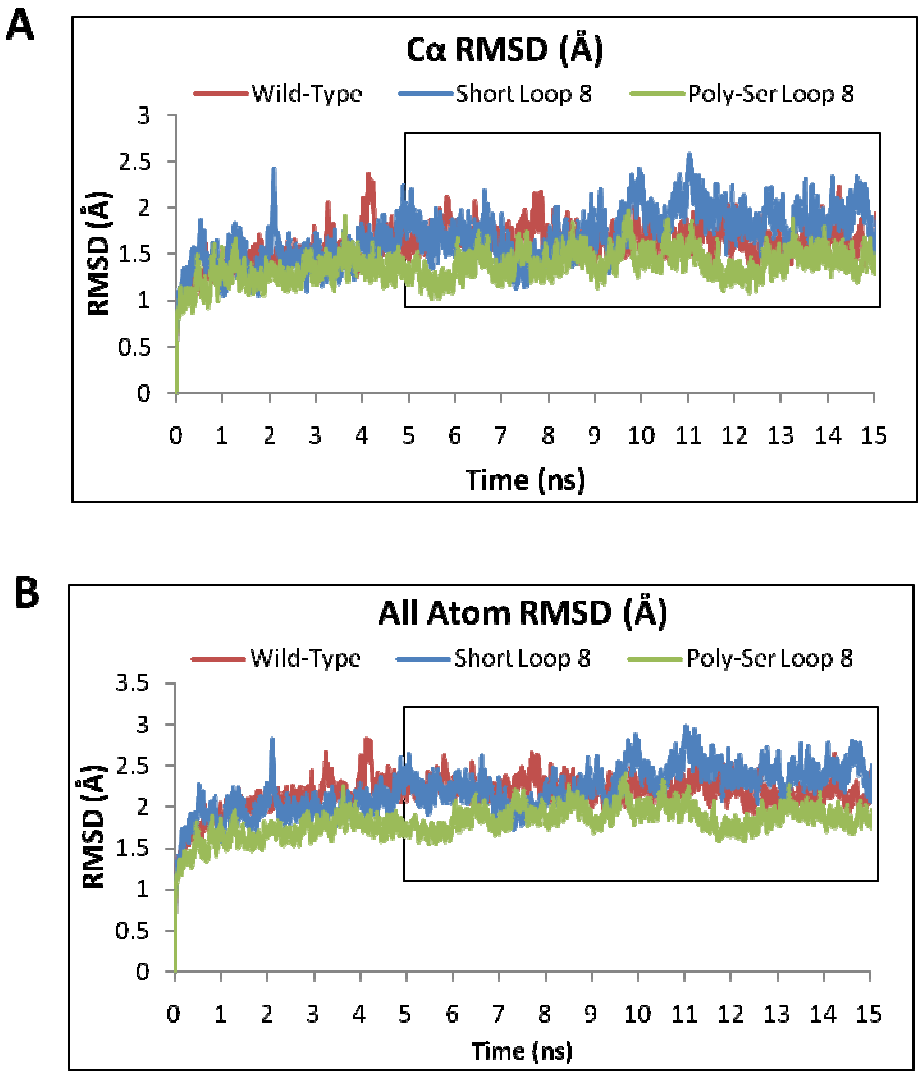


Figure 4.2 Root mean squared deviation of atom positions over the time scale of the MD simulations

(A) Cα RMSD and (**B**) all atoms RMSD over 15 ns molecular dynamics simulation. Boxed regions denote the time span of the simulations used in data analysis. These two figures demonstrate that all three models used in the MD simulations came to equilibrium at the beginning of the production run. It is important to select a constant time frame to compare the three models correctly.

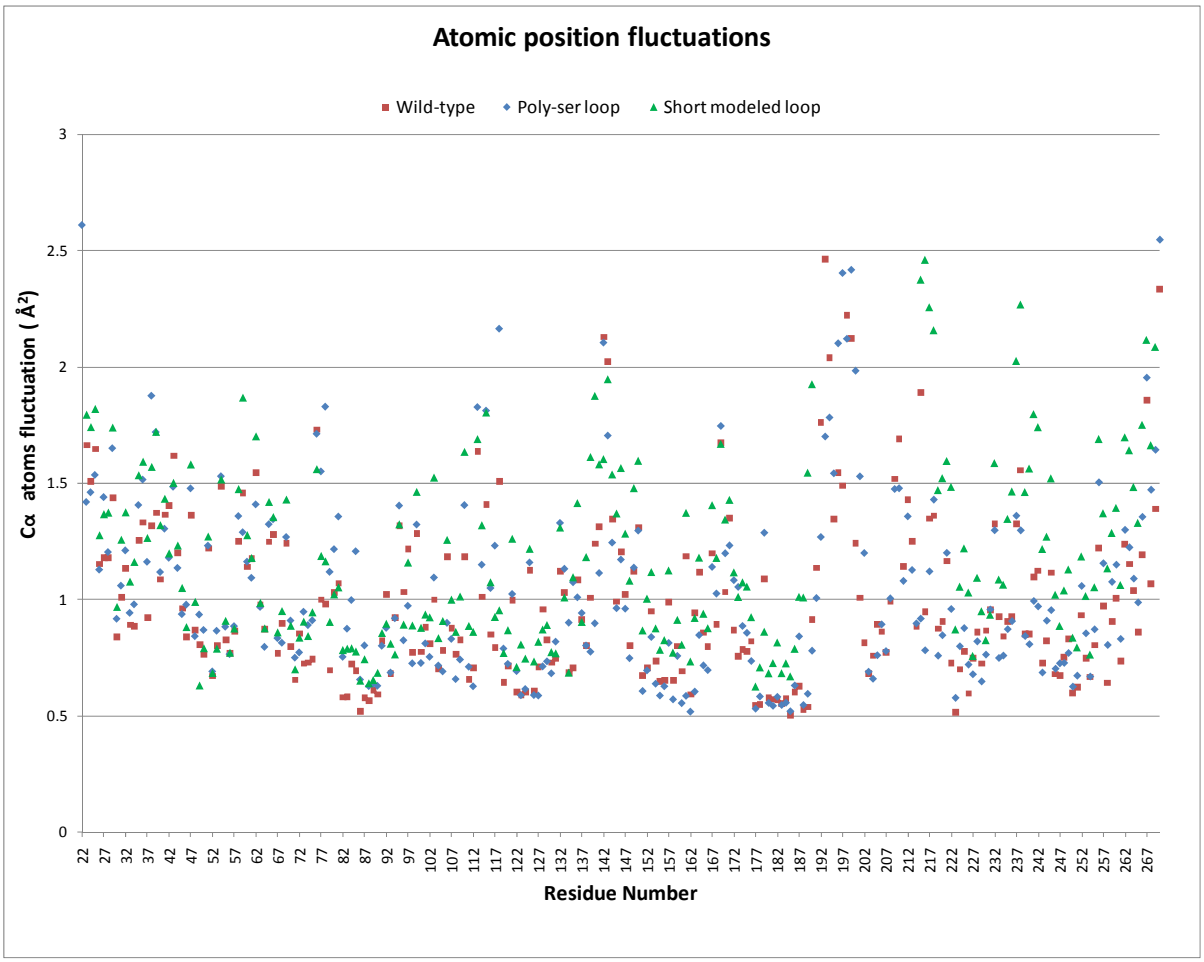


Figure 4.3 Atomic position fluctuations of each Cα position

The x-axis represents the Cα atome of each residue and the y-axis is the atomic position fluctuation of that Cα atom in $Å^2$. The area of highest fluctuation is loop 8 (187-216) in all cases. The majority of helices have lower than 1 $Å^2$ fluctuation. Helical regions are typically less dynamic than loop regions, and that can be seen here.

Table 4.1 Average atomic position fluctuations ($Å^2$) for three replicates of the MD simulations

Cα Atoms	Wild-Type Loop 8	Poly-Ser Loop 8	Short Loop 8
Loop 8 Cα only	1.213±0.035	1.430±0.318	1.682±0.040
All Cα	1.024 ±0.024	1.100±0.087	1.162±0.036
Non Loop 8 Cα	0.998±0.032	1.053±0.056	1.150±0.046

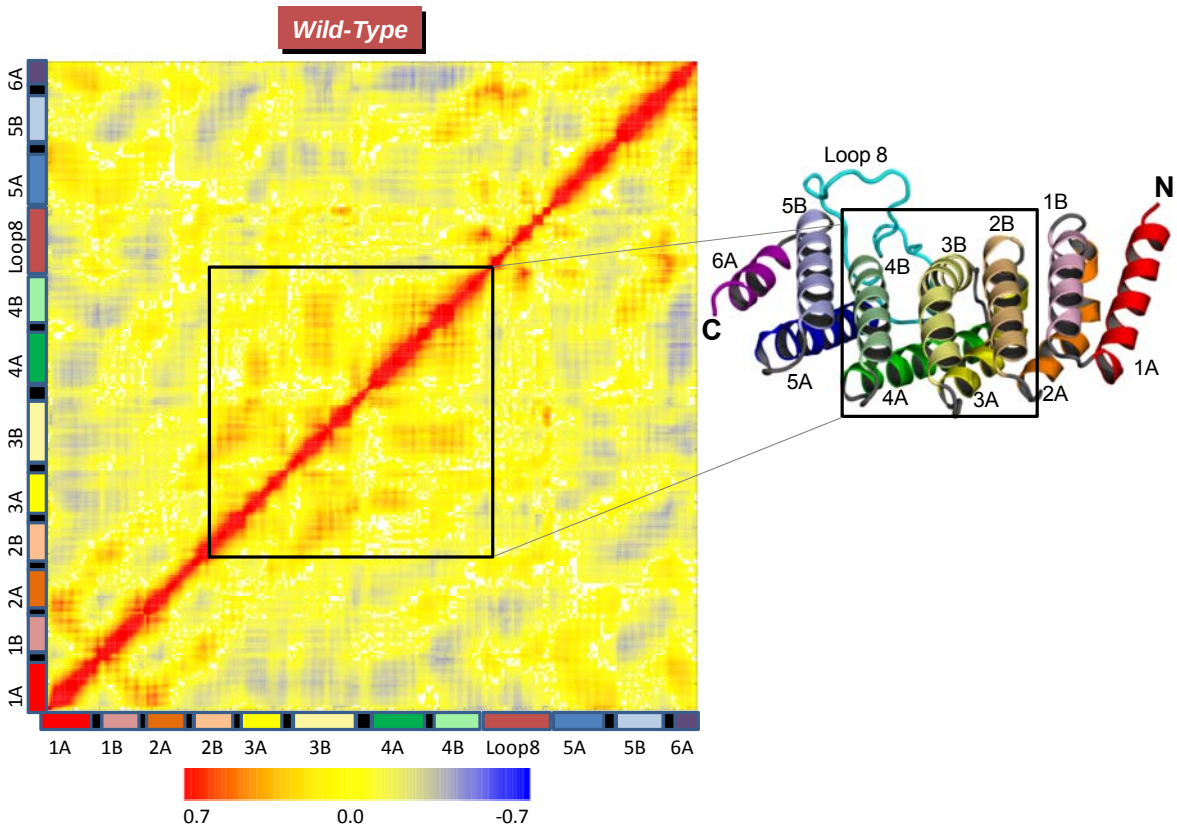


Figure 4.4 Symplekin wild-type correlation plot and structural implications

Correlation plot shows regions of correlated (red) and anti-correlated (blue) motions. The X- and Y-axes represent each residue in the wild-type structure. The bars are colored according to their assigned secondary structural elements as shown in the structure (**Figure 2.1**). The central portion of the plot shows that the three middle HEAT repeats display correlated movement. The scale shows the relative intensity of the correlation. Therefore, regions of yellow still have motion, but are not correlated or anti-correlated movements.

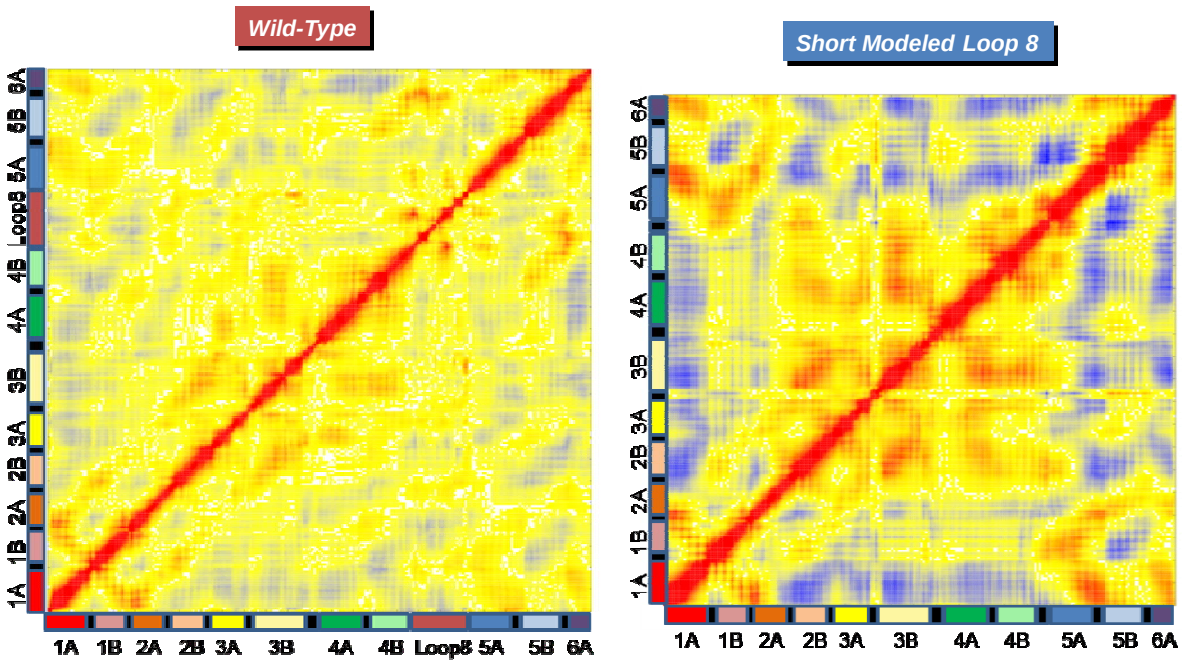


Figure 4.5 Comparison of wild-type and short modeled loop 8 correlation plots

The correlation plots of wild-type and short modeled loop 8. Red areas are correlated and blue areas are anti-correlated. The scale is the same as in Figure 4.4. Again, the central region of the protein, HEAT repeats 2, 3 and 4 move in correlated motion, while the termini move in an anticorrelated motion with the central portion. All motions have greater relationships in model where Loop 8 has been removed and replaced with a loop typically found in HEAT domains. The presence of Loop 8 effects the motion relationships within the protein.

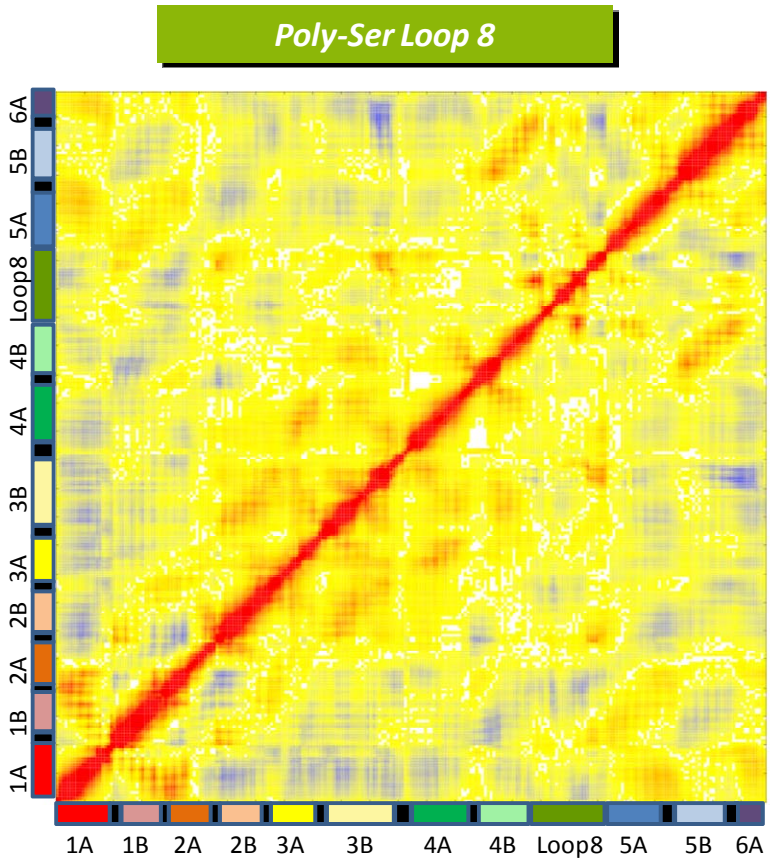


Figure 4.6 Correlation plot of Poly-Ser loop 8 mutant Symplekin

Correlation plot of Poly-Ser loop 8 mutant Symplekin with each residue on the x and y axis. Red indicates correlated movement and blue represents anti-correlated movements. The scale of correlation is the same as in Figure 4.4. This plot is very similar to the level of and arrangement of motions seen in the wild-type MD simulation. It appears that the presence of loop 8 and perhaps not the identity of the residues within the loop are important for motion relationships within the Symplekin HEAT domain.

Table 4.2 List of atoms with electrostatic interactions in the crystal structure

Atom 1				Atom 2			
Name	Residue	Atom name	Atom #	Name	Residue	Atom name	Atom number
LYS	132	1HZ	1745	GLY	200	O	2791
LYS	132	2HZ	1746	GLY	200	O	2791
LYS	132	3HZ	1747	GLY	200	O	2791
SER	203	H	2825	ASP	206	OD1	2872
SER	203	H	2825	ASP	206	OD2	2873
ARG	258	1HH1	3693	SER	203	OG	2831
ARG	258	2HH1	3694	SER	203	OG	2831
ARG	258	1HH2	3696	SER	203	OG	2831
ARG	258	2HH2	3697	SER	203	OG	2831
ARG	258	1HH1	3693	ASP	201	OD1	2800
ARG	258	2HH1	3694	ASP	201	OD1	2800
ARG	258	1HH2	3696	ASP	201	OD1	2800
ARG	258	2HH2	3697	ASP	201	OD1	2800
ARG	258	1HH1	3693	ASP	201	OD2	2801
ARG	258	2HH1	3694	ASP	201	OD2	2801
ARG	258	1HH2	3696	ASP	201	OD2	2801
ARG	258	2HH2	3697	ASP	201	OD2	2801
ARG	258	1HH1	3693	MET	257	O	3675
ARG	258	2HH1	3694	MET	257	O	3675
ARG	258	1HH2	3696	MET	257	O	3675
ARG	258	2HH2	3697	MET	257	O	3675
SER	195	HG	2705	MET	257	O	3675
ASP	192	H	2659	SER	195	OG	2704

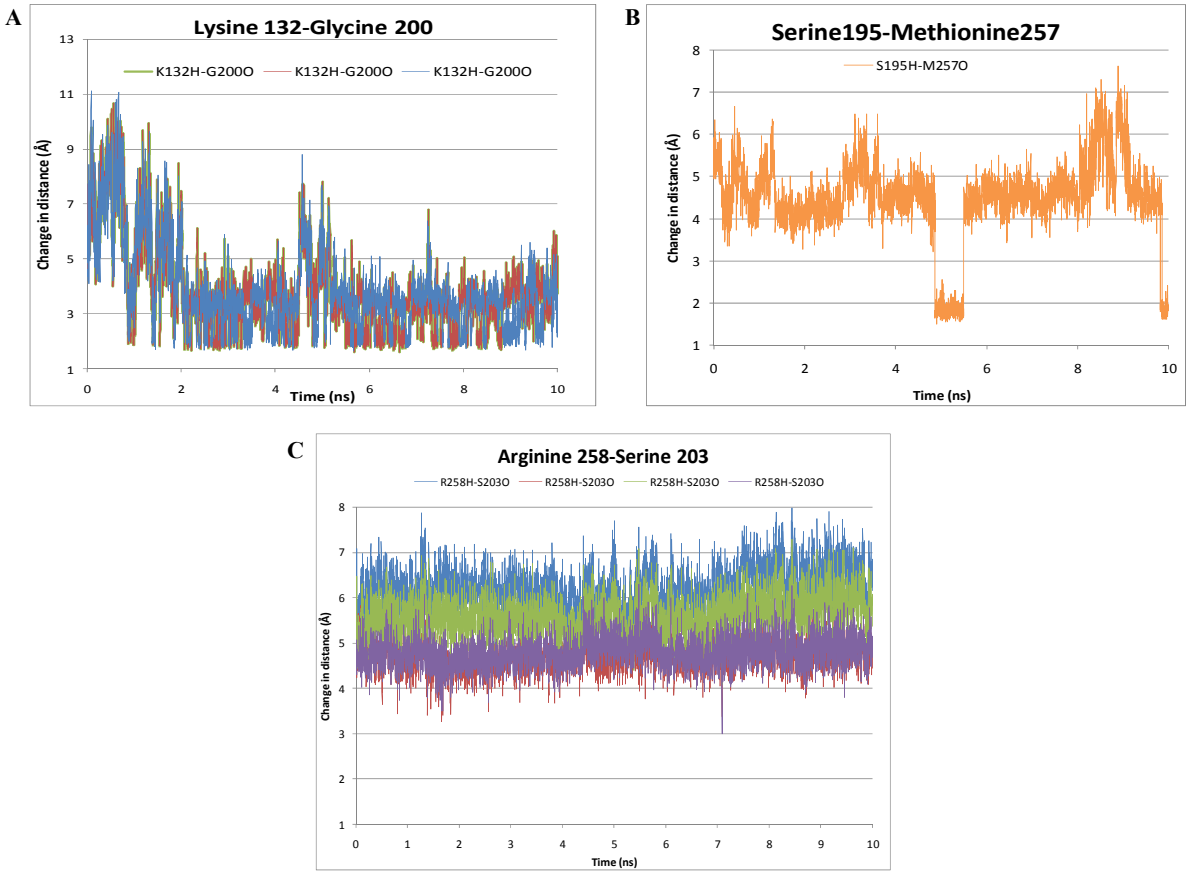


Figure 4.7 Electrostatic interactions disrupted during simulation

Each graph shows the change in distance vs. time, with the specific atoms listed in the legend. (A) Distance between each hydrogen atom attached to the side-chain nitrogen of K132 and the oxygen from the main-chain oxygen of G200. (B) Distance between Hγ of S195 and main-chain Oγ of M257. (C) Four hydrogens on R258 side chains do not maintain electrostatic interaction with Oγ of S203. Since these interactions are disturbed during the simulations, perhaps they are less important for anchoring loop 8 in position at the end of HEAT repeats 3, 4 and 5.

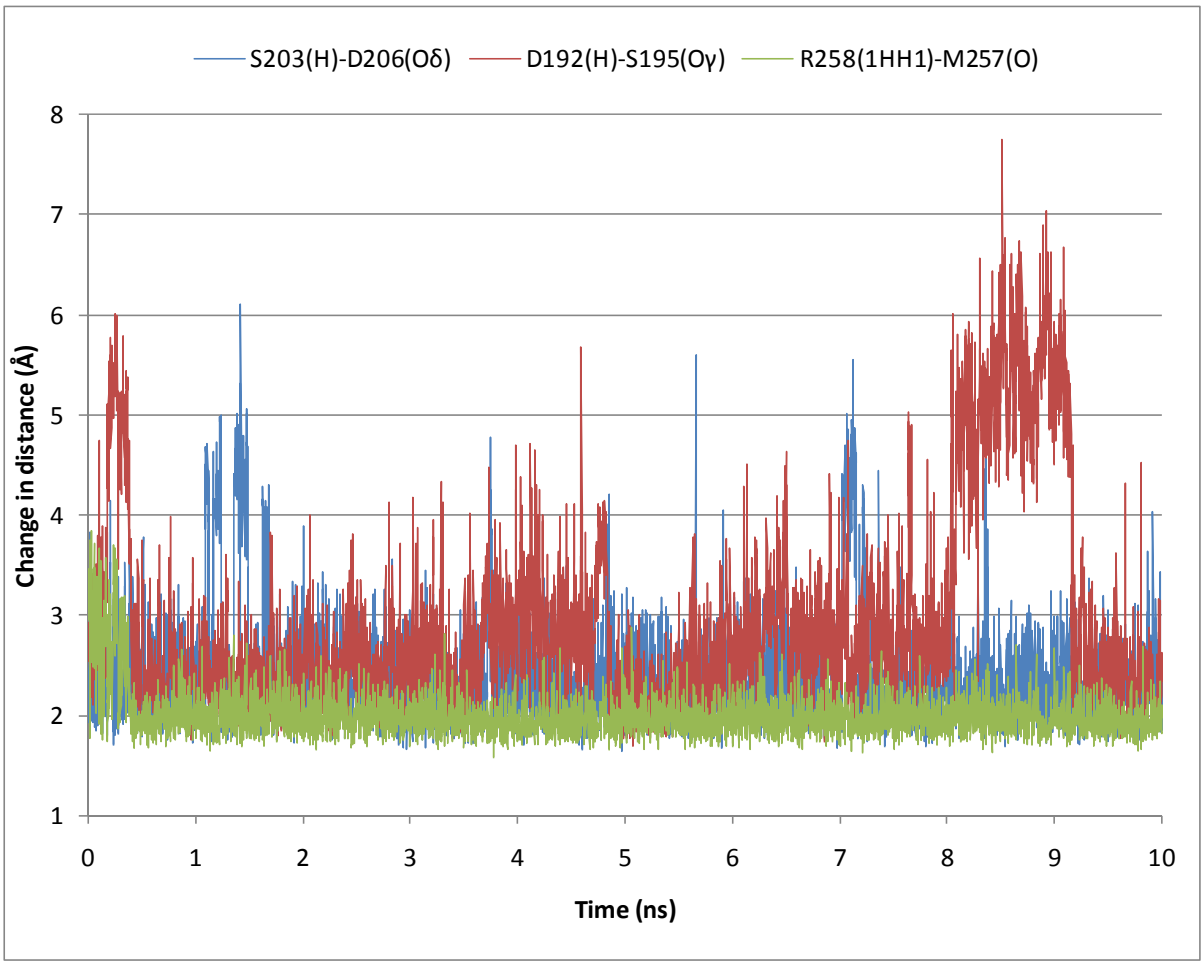


Figure 4.8 Electrostatic interactions maintained during the wild-type simulation

Change in distance vs. time for three electrostatic interactions found in the crystal structure. The distances shown on this graph demonstrate that specific electrostatic interactions remain within close proximity during the wild-type MD simulation. Serine 203 remains within a few angstroms of aspartic acid 206, aspartic acid 192 and serine 195 also maintain close proximity. These sets of residues maintain electrostatic interactions within loop 8 during the 10 ns MD simulation. R258 and M257 also maintain close distance, but this might be due to their proximity in loop 10.

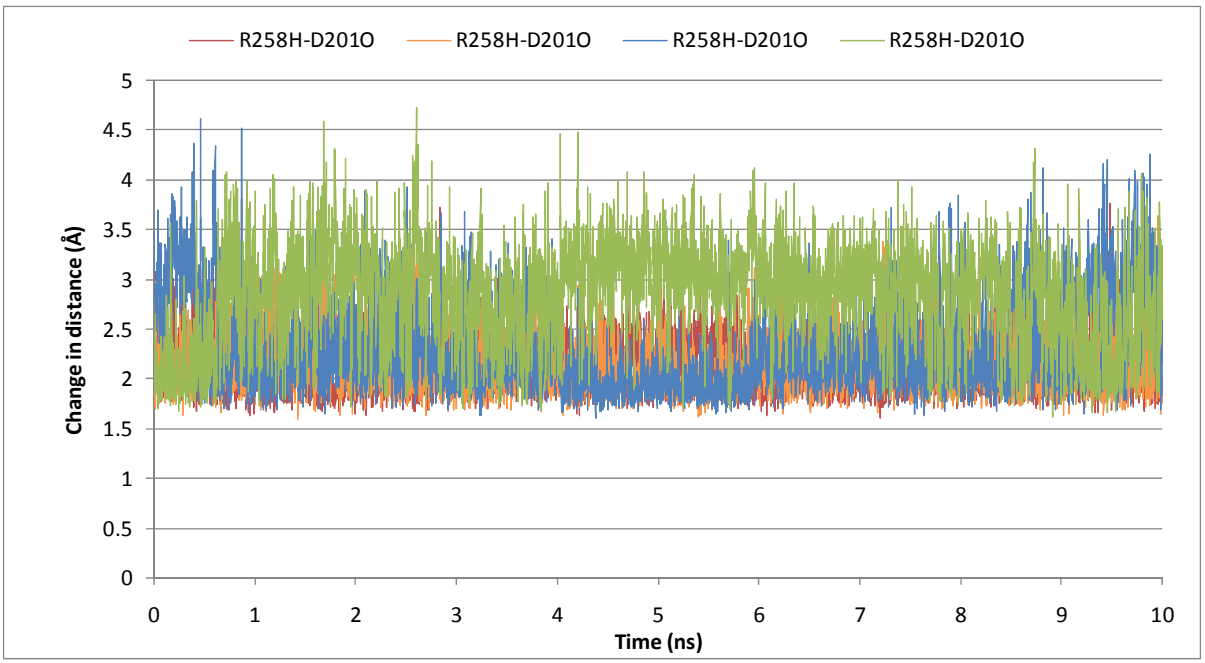


Figure 4.9 Arginine 258 remains in close proximity to aspartic acid 201 during the wild type molecular dynamics simulation

Change in distance vs. time for three electrostatic interactions found in the crystal structure. This key interaction between conserved residues arginine 258 and loop 8 aspartic acid 201 is maintained throughout the MD simulation, as can be seen here by monitoring the distance between hydrogens on the arginine and the two aspartic acid oxygens. Thus, loop 8 may be using D201 to anchor to loop 10 R258.

Chapter 5: Preliminary assays and preparation of Symplekin mutants for biochemical characterization

5.1 Introduction to biochemical characterization of Symplekin

As outlined in chapter one, Symplekin plays a role in 3'-end mRNA processing and has been found at cellular tight junctions. Because Symplekin is involved with processes involving RNA modification and the HEAT domain is predicted to be structurally similar to RNA-binding protein Pumilio, I probed the RNA binding activity of the Symplekin HEAT domain using fluorescence polarization. No interaction with RNA was detected, and all literature on Symplekin thus far has not indicated any RNA binding role.

Symplekin binds numerous proteins as has been shown through pull-down assays, immunofluorescence and immuno-precipitation (**Table 1.1**). Most of these interactions utilize full length Symplekin protein. HSF1, Ssu72 and Glc7p are the only proteins that have been shown to bind specifically to the N-terminus of Symplekin; hHSF1 binds to the first 125 residues of human Symplekin, yeast Ssu72 binds to the first 300 residues of Pta1, while Glc7p interacts with residues 100-200 of Pta1[33,44,45]. To probe binding partners, I worked in collaboration with Dr. Mindy Steiniger of the Marzluff laboratory at UNC-CH on development of pull-down binding assays. Her goal is to work with *D. melanogaster* proteins and reconstitute the full length and HEAT domain Symplekin interactions with homologues of HSF1, Ssu72, CPSF73 and CstF-64. Mindy is currently developing the protein-binding interactions through pull-down assays and histone processing reactions; I show some of our initial results in this chapter. Finally, at the end of the chapter, I discuss

the preparation of several mutations within the Symplekin HEAT domain for testing in the finalized protein binding assay.

5.2 Initial nucleic acid binding assays for Symplekin are inconclusive

SVM-prot, an online tool that identifies protein function, indicated that Symplekin may bind RNA, predicted with an R-value of 1.4 and P value of 71.3%[91]. Symplekin has striking similarity to the RNA-binding Pumilio family of proteins: Z-score of 9.5 over 160 aligned residues with an RMSD of 3.1 Å[75,79] (**Figure 5.1**). Fluorescence polarization assays were performed to test the hypothesis that Symplekin's HEAT domain binds to RNA. Assays were developed based on the fluorescence anisotropy experiments of LeTilly *et al.* and those of Aviv *et al*[92,93].

Examination of the alignment indicates that the specific Pumilio residues participating in RNA-binding are not conserved in Symplekin, so it is unlikely that they would bind to the same RNA sequence. 5'-Fluorescein-labeled RNA substrates including a 10 nucleotide sequence mimicking the Pumilio-binding RNA sequence, a 29 nucleotide stem-loop sequence conserved within histone mRNA, and a random 30 nucleotide sequence (also containing a stem-loop) were utilized in an fluorescence polarization binding assay with Symplekin 19-271 (**Table 5.1**). Fluorescein absorbs at 485 nm and emits at 520 nm, so a change in anisotropy can be detected by monitoring the change in absorption and emission at these wavelengths. A 5'-fluorescein-labeled 15-mer DNA strand was used in place of the RNA as a negative control reaction.

During the assay, the concentration of nucleic acid was held constant at 10 or 100 nM, while the protein concentration ranged from 150 μM to 0.001 μM. The Symplekin

HEAT domain protein was purified as outlined in chapter 2. The buffer for the binding assay was 20 mM HEPES, pH 7.4, 0.1 mM NaCl, 0.1mM EDTA and 20% glycerol. The buffer, protein and substrate were added to a 384-well plate and gently spun down to remove any air bubbles. The solutions were allowed to sit for 2 minutes and then the anisotropy measurements were read using the BMG Pherastar.

Figure 5.2A displays a preliminary binding assay to determine the ideal range of protein concentrations to examine. The substrates (15-mer DNA and the mmu RNA, see **Table 5.1**) in this figure are held at 10 nM. No significant change in anisotropy is seen until Symplekin protein levels reach 0.25 μM. Also, it appeared that perhaps there was a difference between the RNA and the DNA binding profiles at higher protein concentrations; however change in anisotropy was very minimal. Another assay was performed to validate this result, utilizing higher concentrations of protein. To ensure enough time for binding, during the second round of assays, the nucleic acid and protein were combined in the plate and allowed to incubate for 30 minutes before testing. **Figure 5.2B** shows that there is no binding of Symplekin HEAT domain to any substrate. A proper binding curve would show an increase in fluorescence units as the Symplekin HEAT domain protein concentration is increased. The results are erratic in each run, indicating the results are not corresponding to a binding event. Also, there is no positive control reaction to validate the assay, so no conclusions can be drawn from these initial experiments. The nucleic acid substrates and proteins are available for further experiments and the Pumilio protein can be used as a positive control protein.

The RNA-binding function predicted by SVM-prot was no doubt based on the fact that Symplekin and Pumilio contain similar structural motifs. Throughout my literature

search and structural alignment work, the majority of proteins with HEAT repeats bind to proteins, not nucleic acids. Also, the interactions between the conserved sequences in the 3'-end of messenger RNA already have well documented protein binding partners. This underwhelming RNA binding data, taken together with known protein-protein interactions between Symplekin and CstF64, CstF77, CPSF73, ZONAB,HSF-1, Ssu72 and many others (**Table 1.1**) indicate that the Symplekin is poised for protein binding and not RNA binding.

5.3 Preliminary pull-down assays to characterize Symplekin interactions with its putative protein binding partners

The GATEWAY system was utilized for cloning Symplekin's multiple putative protein binding partners. This system from Invitrogen utilizes a pENTR vector that is useful to shuttle the gene to different functional plasmids. Thus, the versatility of the pENTR vector will be advantageous for studying recombinant proteins and *in vivo* experiments for investigation of mRNA processing reactions. For preparation of recombinant protein interaction assays, the pENTR clones were shuttled to the pEXP2 vector that provides a C-terminal 6xHis tag. **Table 5.2** shows the list of constructs that are in various stages of cloning into the pENTR vector and shuttled to the pEXP2 vector. Original cDNA for cloning was supplied by Deirdre Tatomer in the Biology Department of UNC-CH and the cloning and expression were done in collaboration with Dr. Mindy Steiniger in Dr. William Marzluff's laboratory at UNC-CH.

Since these proteins are all of eukaryotic origin, they were expressed in an *in vitro* translation system that couples transcription and translation (TnT, Promega). This cell-free method utilizes crude extracts from rabbit reticulocyte lysate that contain all the components

required for translation and a prokaryotic phage RNA polymerase for transcription from plasmid DNA. Radio-labeled amino acids were utilized in the media as a method to detect protein expression because protein yield is only a few nanograms per 50 uL reaction. **Figure 5.3** illustrates the TnT expression of three CstF64 constructs and full length ssu72.

In collaboration with Dr. Mindy Steiniger, pull-down assays were carried out to assess the ability of the HEAT domain to interact with the TnT expressed proteins. Recombinant Symplekin HEAT domain tagged with N-terminal 6xhis-MBP was purified (see chapter 2) and added to amylose resin for MBP binding. CstF64 and Ssu72 proteins prepared through TnT were added to separate batches of recombinant MBP-Symplekin bound to amylose. Proteins were incubated for at least 30 minutes in PBS buffer. Reactions were then batch washed 3 times and eluted with SDS or PBS-maltose. Unfortunately, both Symplekin HEAT domain and MBP-only controls showed interactions with ssu72 and CstF64 constructs (**Figure 5.4A**). More salt was added to the binding and wash buffer in an attempt to reduce the non-specific MBP protein binding to the TnT expressed protien. However, the same level of binding was seen with MBP or MBP-Symp 19-271 even with up to 300 mM NaCl added to the binding buffer (**Figure 5.4B**). This amylose-MBP pull-down was an ineffective method to examine interactions between Symplekin 19-271 and ssu72 or CstF64 constructs. Utilizing a method that does not include MBP or amylose was the next course of action.

The next round of binding assays were attempted using full length N-terminal GST tagged Symplekin. This time, Symplekin was bound to GST resin and the pull-downs were performed in a similar manner, except proteins were incubated for several hours in PBS buffer, PBS buffer plus 1M NaCl, or PBS buffer plus 0.1% CHAPS. The protein was eluted

with PBS plus glutathione. During this round, CPSF73, full length CstF64 and CstF64-hinge were tested for binding to full-length Symplekin (**Figure 5.5**). Symplekin and CstF64 were shown to co-immunoprecipitate in human cells, so this interaction was thought to be a positive control to validate this assay[26]. Again, the GST only negative control reaction also shows interaction with the TnT expressed proteins. Unfortunately, up to this point, we have been unable to positively identify any protein binding partners for Symplekin through the pull down assays.

More effort is needed to elucidate protein-protein interactions of Symplekin within the mRNA processing machinery. The assays must be optimized and have proper positive and negative controls working before any useful information will be gleaned from the experiments. This is one of Dr. Mindy Steiniger goals at this time; she has submitted a supplemental grant with Dr. William Marzluff to secure the future funding for this part of the project.

5.4 Symplekin HEAT domain mutations for biochemical analysis

Anticipating the development of functional biochemical readout, many mutations were made in the Symplekin HEAT domain. The most striking feature about the structure is the lengthy loop 8 that connects HEAT repeats 4 and 5. To understand if this loop is important for protein binding, four different loop mutations were designed based on the characteristics of the native loop and characteristics of other loops in HEAT repeat units. Each of these mutations simultaneously removed residues alanine 191 through arginine 215, while adding in non-native residues, using one step site-directed mutagenesis reactions. First, a commonly used -G-S-G- loop was inserted between residues F190 and R216. Second, a

single alanine was placed in this loop to see how restricted movement would affect function. Third, in an effort to keep a polar, uncharged residue resting above W152 and Y131, a -G-G-Y-A- linker was inserted. Finally, to keep a hydrophobic cap on these two residues, -G-S-F-G- were added in place of loop 8 residues 191-215. Each of these mutations has been made in the LIC MBP plasmid in Symplekin 19-271.

Deleting the whole loop is one approach to understanding the loop function, but we also wanted to see if the position of the loop over the concave surface seen in the crystal structure is important for function. Thus, the anchoring residues (**Figure 2.3**) were mutated using site-directed mutagenesis. In loop 8, aspartic acid 201, serine 203, proline 208, were all mutated separately to alanine. Residues lysine 132 and arginine 258, on the concave surface of the protein, were also mutated because they are in polar contact with loop 8 to form a pocket between the loop and helices 3B, 4B and 5B. The biochemical characteristics of aspartic acid 201 and arginine 258 mutations will be very interesting because their interaction was maintained throughout the wild-type molecular dynamics simulation (**Figure 4.9**).

The above mutated DNAs were made and sequenced, but the proteins were not all expressed and purified. The sequenced plasmids have a yellow sticker on the top of the eppindorf tube denoting the construct and mutant residue(s). (Note: any eppendorf tube in my plasmid stocks with a sticker on the lid has been sequence verified in the region of the mutation.) When a testable assay is developed by our collaborators, these mutations can be utilized to probe the function of loop 8 and the pocket formed by loop 8 and the concave helices 3-5B.

5.5 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 5.

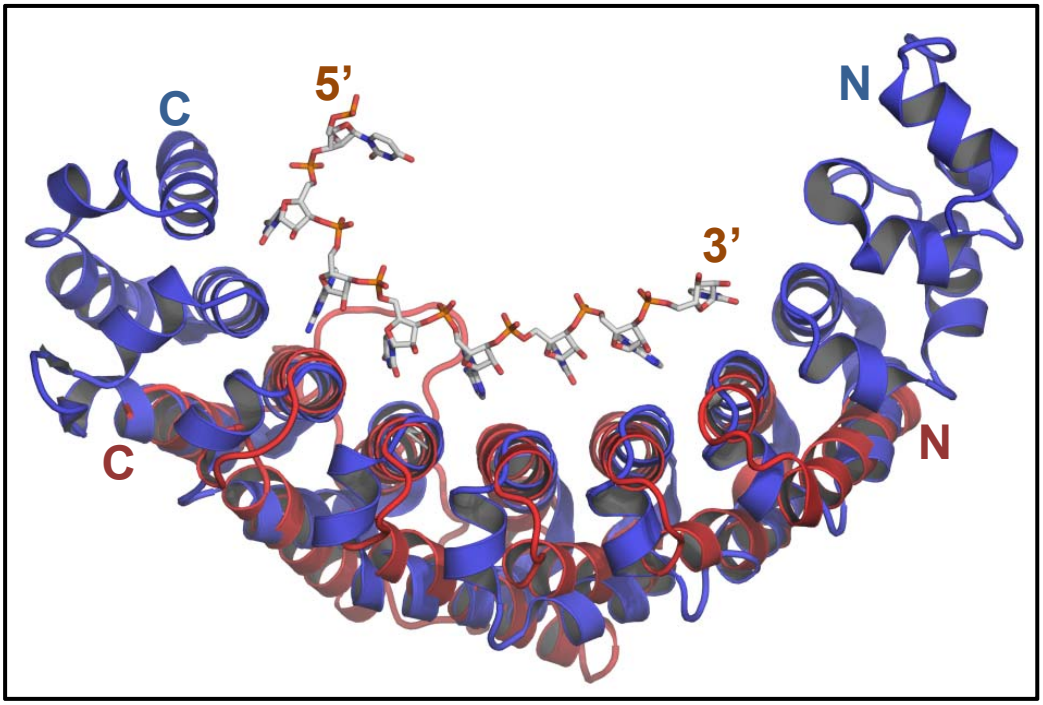
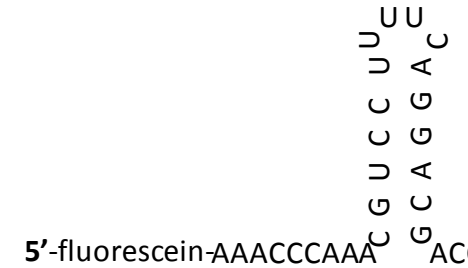


Figure 5.1 Symplekin HEAT domain structurally aligned with Pumilio bound to RNA

Pumilio (PDB: 1m8y), blue, uses its concave face to bind the bases of the RNA strand. Symplekin, red, has been structurally aligned to Pumilio using Dali[74].

Table 5.1 Nucleic acid substrates for Symplekin HEAT domain binding assays

NAME	SEQUENCE/STRUCTURE
Puf-binding	**5'**-fluorescein-CUUGUACAUA **3'**
Histone stem-loop	
mmu-let-7d	
15-mer ss DNA	**5'**-fluorescein-ACAAGGAGGTCGTCA **3'**

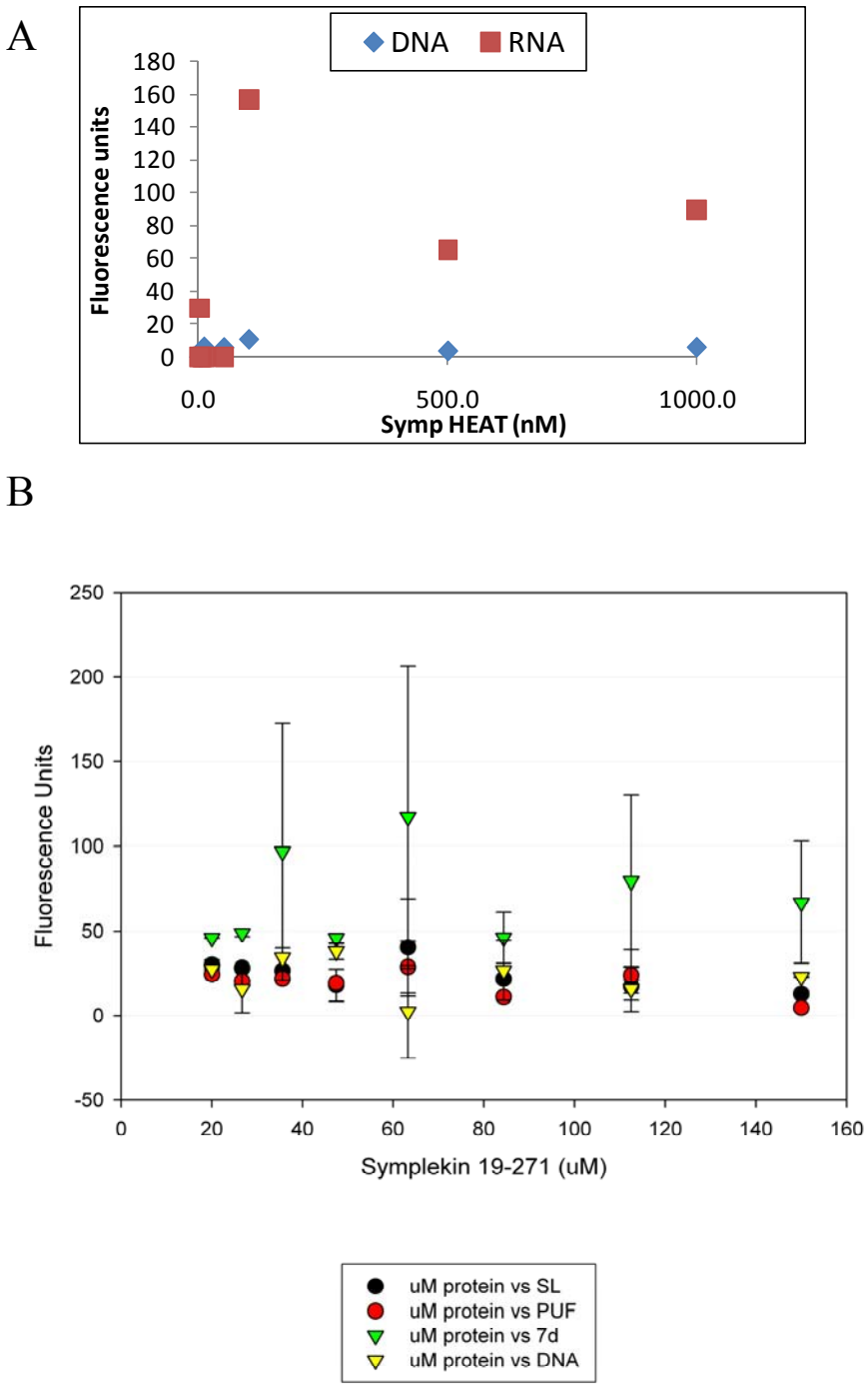


Figure 5.2 Nucleic acid binding assays with Symplekin HEAT domain

(A) Initial binding assay with 15-mer ssDNA and mmu-let-7d RNA. (B) Binding assay with Symplekin HEAT domain and each of the four substrates. No conclusion can be drawn from this work about Symplekin binding to DNA or RNA. Further work must be done to elucidate this role for Symplekin.

Table 5.2 List of proteins to be assayed for binding Symplekin

Name	Fragment	Fragment name	Cloned into pENTR	cloned into pEXP-2
CstF 64	272-419	structured region	x	
	1-419	full length	x	x
	108-214	hinge	x	x
	215-419	C-term, Δhinge	x	x
YPS	1-352	full length	x	
ssu72	1-195	full length	x	x
Symplekin	1-1165	full length	x	x
CPSF100	1-271	HEAT domain	x	x
	1-756	full length	x	x
	1-524	β-lactamase/β-casp		
	525-756	putative dimerization		
CPSF73	1-684	full length	x	x
	1-465	β-lactamase/β-casp		
	466-684	putative dimerization	x	x
HSF1	1-691	full length		
CG8816	1-175	full length		

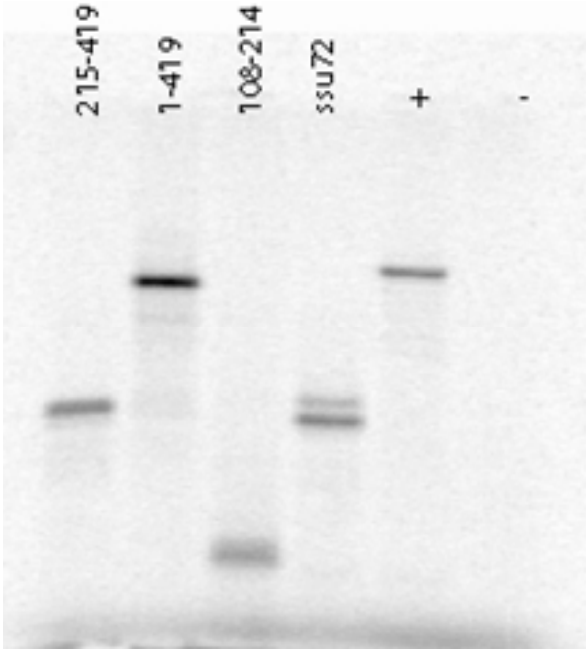


Figure 5.3 TnT expression of CstF 64 and ssu72 on SDS-gel visualized by radiography

CstF64 C-terminus (215-419), full length (1-419), hinge region (108-214) and full length ssu72 were expressed through *in vitro* translation with ^{35}S-radiolabeled amino acids to visualize the protein by radiography. Positive and negative controls were run to ensure translation. Positive control protein was Lsm11, while the negative control was empty vector.

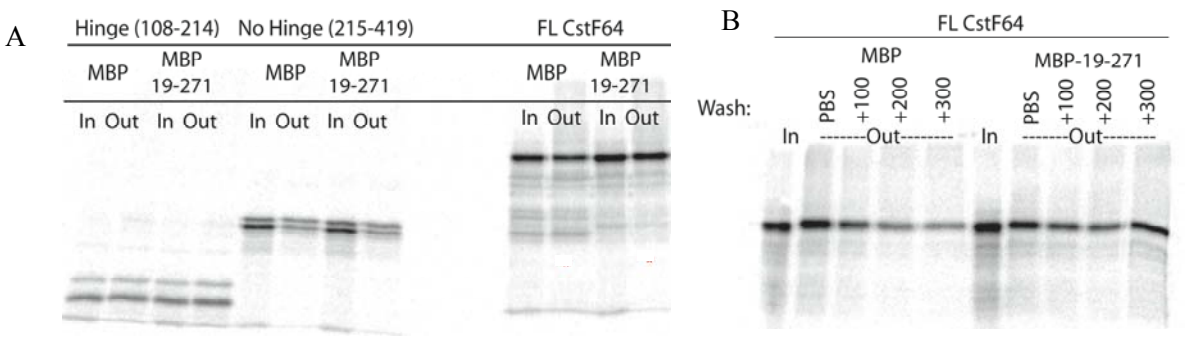


Figure 5.4 Pull-down experiments with CstF64 and the Symplekin HEAT domain using amylose resin

(A) CstF64 fragments assayed for binding with MBP (negative control) or MBP-Symplekin 19-271. "In" represents 10% of the total input. (B) Full length CstF64 assayed for binding to MBP (negative control) or MBP-Symplekin. Assay buffer conditions are listed as PBS or PBS +100, +200 or +300, indicating the amount of NaCl added to the PBS buffer.

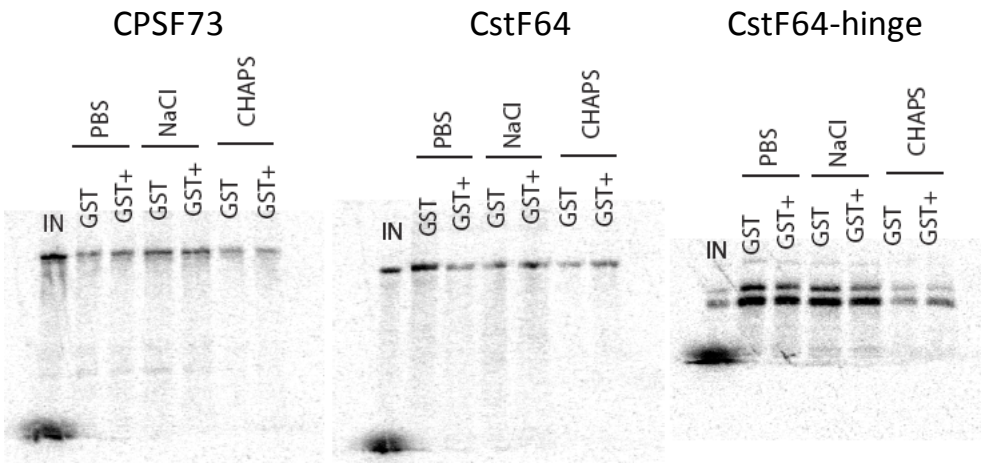


Figure 5.5 Pull-down experiments with CPSF 73, CstF 64 and CstF64-hinge binding to full length Symplekin using GST affinity resin

N-terminal GST tagged full length Symplekin or GST alone (negative control) assayed for binding with radio-labeled full length CPSF73, CstF64 or CstF-hinge region. Three binding buffers were examined: PBS, PBS + 1M NaCl, and PBS + 0.1% CHAPS. GST+ stands for GST-Symplekin. As illustrated, binding was seen with GST alone, and therefore the results of GST-Symplekin binding are inconclusive.

Chapter 6: A novel fold in the TraI relaxase-helicase C-terminus is essential for conjugative DNA transfer

6.1 Introduction to TraI and its role in conjugative DNA transfer

Conjugative DNA transfer (CDT) is one way that cells can transfer DNA from donor to recipient. This process of horizontal gene transfer occurs between bacterial cells, plant cells and across these kingdoms[94-97]. One of the main reasons that this area of study has been so prevalent is because CDT is a way to spread antibiotic resistance genes on mobile plasmids, thus creating new strains of drug-resistance microbes[98,99]. In fact, the centers for disease control and prevention has made an action plan to address the escalating crisis of antibiotic resistant bacteria, and scientists in academic, industrial and government settings are fighting this growing epidemic. To fight this problem, we must understand the mechanism and individual components of the bacterial conjugation. CDT requires a close cell-cell junction through which the genetic material is transferred. In the well-studied model F plasmid system in *E. coli*, CDT requires the type-IV secretion system (T4SS or TFSS) and a 500 kD relaxosome. The overall process is outlined in **Figure 6.1**. A cell with the F plasmid (F^+) assembles the type-IV secretion system to initiate cell contact with a recipient (F^-) cell. Then, the relaxosome nicks and unwinds a single-strand of the F-plasmid and transfers the DNA to the recipient cell. This yields a new F^+ cell.

Figure 6.2 illustrates the proteins composing the relaxosome in the F-plasmid system. TraM, TraY and TraI are all encoded in the F-plasmid, while IHF (integration host factor) is encoded in the bacterial host genome. These proteins have specific DNA binding sites on the

origin of transfer (oriT) in the F-plasmid[100-103]. IHF bends the plasmid DNA[104], which may bring TraY and TraM in closer contact with TraI. TraY and TraM functions are not fully understood, however TraY and IHF must be present before TraI is able to bind to the *oriT nic* site[105]. TraI is the main catalytic component of the relaxosome and functions to both nick and unwind the DNA for transfer[106].

TraI is a multi-domain protein, consisting of an N-terminal relaxase, a central region of unknown function followed by a helicase domain and a C-terminal domain (**Figure 6.3**). The N-terminal relaxase contains highly conserved tyrosine residues and an HUH motif that are responsible for coordinating a transesterification reaction, cleaving the 5' side of the DNA phosphate and creating a free 3'-hydroxyl; this domain binds site-specific ssDNA with a sub-nanomolar K_D[107]. Structures of this domain have been solved by the Redinbo and Schildbach groups[108,109]. The helicase domain encompasses an ATPase domain, which is responsible for energetically driving the unwinding of the F-plasmid[110]. While the detailed role of the C-terminus region is unknown, Matson *et al.* provide indirect evidence that indicates deletion of this region disrupts conjugative DNA transfer[111]. It has also been speculated that the TraI C-terminus binds to TraM to stimulate the transesterification reaction *in vivo*[112].

The Redinbo group has focused on understanding the structure and function of the TraI protein. Scott Lujan solved the structure of the F-plasmid TraI relaxase domain and worked on designing inhibitors to this domain, to decrease plasmid transfer[109]. Dr. Redinbo has started a company to commercially develop inhibitors of TraI to prevent spread of antibiotic resistance genes. Two other graduate students are studying the relaxase domains in other plasmid systems and one student is investigating the F-plasmid TraI helicase domain.

Dr. Laura Guogas and I collaborated to structurally and functionally characterize the C-terminal domain of F-plasmid TraI; the following sections describe the work we did to characterize this C-terminal domain. Specifically, Dr. Guogas did the initial structure determination, and worked on the DNA binding assays and conjugation assays with Jin-Hyup Lee, while I completed the structure and utilized dynamic light scattering (DLS) to study the oligomeric state of the protein. Our paper, "A Novel Fold in the TraI Relaxase-Helicase C-Terminal Domain is Essential for Conjugative DNA Transfer" was accepted to the *Journal of Molecular Biology* in December 2008[113].

6.2 Sequence conservation in the 1476-1629 TraI C-terminal region

TraI orthologs in related *Yersinia pestis, Klebsiella pneumoniae, Salmonella typhi, Aeromonas salmonicida, Enterobacter* sp. 638, *Shigella sonnei,* and the *E. coli* R-100 conjugative plasmids were compared to the F plasmid TraI sequence using ClustalX[114]. A high degree of conservation (27-99% sequence identity) is maintained within the 1476-1629 region ordered in our crystal structure. In contrast, significant sequence divergence is observed in the far C-terminal region (1630-1756) (**Figure 6.4**). Only the *E. coli* R-100 resistance plasmid and a plasmid from *S. sonnei* maintain high sequence identity (96%) through the far C-terminus of their TraI proteins. In contrast, the other plasmids exhibit only 13-44% sequence identify within this region. Thus, residues 1476-1629 in the TraI-CT comprise a conserved core domain present in TraI orthologs in a range of F-like conjugative plasmids.

6.3 Solving the TraI C-terminal structure

The gene encoding the TraI C-terminal domain (residues 1476-1756) was cloned into the vector pMCGS9[56] that fuses the expressed protein C-terminal to maltose binding protein (MBP) and a 6-His tag. A TEV cleavage site is located between the target protein and the tags. One liter flasks of LB broth were inoculated at a ratio of 1:100 from a saturated overnight culture of BL21 cells containing this protein-expression plasmid. Cells were grown at 37 °C under antibiotic selection with vigorous shaking until the cell density reached an OD_{600} of 0.6. IPTG was then added to a final concentration of 0.5 mM and the temperature was dropped to 16 °C for overnight expression. Cells were pelleted and resuspended in Nickel A buffer (20 mM Tris-HCl pH 7.4, 5% glycerol, 20 mM imidazole and 300 mM NaCl) at a ratio of 10 mL buffer/L of original culture. Lysis was carried out using a sonicator (Heat Systems) pulsed on ice for approximately three minutes. Following centrifugation at 27,000 g, the cleared lysate was loaded onto a gravity column packed with Nickel Sepharose 6 Fast Flow resin (Amersham) pre-equilibrated with Nickel A Buffer. Following washing to baseline, the His-tagged protein was eluted with a high imidazole Nickel B buffer (20 mM Tris-Hcl pH 7.4, 5% glycerol, 500 mM imidazole and 300 mM NaCl). Fractions containing the TraI-CT MBP fusion protein (as evaluated by SDS-PAGE) were pooled and cleaved with 1% (w/w) TEV protease during dialysis into Nickel A buffer overnight at 4 °C. Following cleavage, a second run through the nickel column separated TraI-CT from MBP; TraI-CT flowed through the column while the His-tagged MBP remained bound to the column and was later eluted with Nickel B buffer. The purity of TraI-CT was assessed by SDS gel electrophoresis. Fractions containing clean TraI- CT were pooled for a final polishing step. The protein was loaded onto a 26/60 Superdex 75 size exclusion column in 20 mM HEPES-KOH pH 7.4, 150 mM NaCl on an Akta Express FPLC

(GE Healthcare). TraI-CT peak fractions were concentrated and buffer exchanged into 150 mM ammonium acetate using an Amicon ultracentrifugation filter (Millipore). TraI-CT was concentrated to 25 mg/ml as determined by $UV_{280\ nm}$ measurement.

TraI-CT containing selenomethionine was generated using B834 cells, a methionine auxotroph cell line. Cells were grown in selenomethionine specific media (Athena) supplemented with 50 mg/L selenomethionine. Expression and purification were preformed as described above. Additional proteins (TraM, TraI full length, TraI 1476-1630, and TraI 1630-1756) were expressed and purified using the same protocol.

TraI-CT and selenomethionyl TraI-CT crystals were grown by hanging drop vapor diffusion at 25 °C. Equal volumes of TraI-CT protein solution and well solution (1 M ammonium sulfate, 100 mM MES pH 6.0) were mixed, and crystals appeared in 7-10 days. Data were collected from a single crystal cryoprotected in 1.8 M lithium sulfate and flash cooled in liquid nitrogen. A three wavelength MAD data set was collected at Sector 22-BM (SER-CAT) of the Advanced Photon Source, Argonne National Laboratory. Data were indexed and scaled using HKL2000[115] (**Table 6.1,** initial solution).

Dr. Guogas originally solved the structure in space group $C222_1$ to 2.1 Å resolution. Data statistics for this solution are found in **Table 6.1**. However, after many rounds of refinement, the R factors remained above the acceptable level for 2.1 Å resolution (typically, R factors should be 10% of the resolution) and several other statistics were amiss: the completeness in the highest shell was only 60%, and the I/σ was less than 3. Thus, Dr. Guogas thought perhaps the wrong space group was assigned, and so she performed molecular replacement with the model from the $C222_1$ dataset and processed her original data into the space group $P2_1$. However, this solution was still unsatisfactory.

By completely reprocessing the data, I was able to discern the reason for the trapped R factors. The data was originally extended to 2.1 Å, and the data in the shells between 2.4-2.1 Å did not exhibit good quality statistics and therefore poor data had been incorporated into the structure solution. **Table 6.2** shows statistics for reflections in the highest 2 resolution shells and total reflections from processing the MAD dataset in the $C222_1$ space group. The overall average redundancy and completeness are higher in the data that has been cut back to 2.4 Å. For example, the highest resolution shell of the 2.10 Å data has only 63.6% completeness. The highest resolution shell in the 2.4 Å data has 89.4% completeness, a value equivalent to the overall completeness for all reflections in the 2.10 Å data (89.3%). The overall completeness for the data cut back to 2.4 Å has a much more acceptable value of 97.5%. Considering the intensity and error in the data, the two highest resolution shells for the 2.1 Å data have 53.1% and 57.2% of reflections with an I/σ value <2, respectively (I is intensity and σ is the error in the intensity). In other words, more than half of the reflections in the highest resolution shells have intensities that are only 2x greater than the error. Trimming reflections to 2.4 Å reduced I/σ in the highest shell to 46.8%. Including all reflections, I/σ is < 2 in 31.1% of the 2.1 Å data, but only 23.4% in the 2.4 Å data. Thus, the highest accepted resolution for the final structure is 2.4 Å and I resolved the structure with the new resolution limit.

Four of ten possible selenium atoms in the two molecules in the asymmetric unit (five methionines per monomer) were located by the program SHARP[116] and used to build an initial model. Two methionine residues (1479, 1588) per monomer were located in the ordered 1476-1629 region of the structure, while the remaining three methionines (1672, 1739, and 1743) per monomer were located in the missing 1630-1756 region. The electron

density map was improved by a combination of solvent flipping using the program SOLOMON[117] and density modification using DM[118], both operating under SHARP. Automatic model building using ARP/wARP placed approximately 100 alanine residues in each monomer. The remaining residues and side chains were placed manually using the program Coot[119]. Model refinement was conducted using the maximum likelihood method in REFMAC 5.2[120], CNS[121], and the free R-factor (with 5% of the data set aside for free R). The final R/R_{free} factors of 0.22/0.26 were reached after several rounds of refinement (**Table 6.3**). The final asymmetric unit contains two protein monomers forming a dimer, along with 130 water molecules and 10 sulfate ions. The C-terminal residues 1630-1756 of the protein construct were missing in each protein monomer and are not present in the final refined model.

The experimental maps produced following phasing and solvent flattening (**Figure 6.5**) allowed the building of a dimer of residues 1476-1629, approximately two-thirds, of the TraI-CT in the asymmetric unit. Residues in the far C-terminus (1630-1756) could not be placed in the model. The Matthew's parameter (V_M) for two molecules of the 1476-1629 region in the asymmetric unit was 2.8; in contrast, two molecules of the complete TraI C-terminal domain (residues 1476-1756) would exhibit a V_M of 1.5. Thus, it is unlikely that the far C-terminal region is present in the crystal. Indeed, SDS-PAGE of washed and dissolved crystal specimen supports this conclusion (data not shown). It is possible that a peptide cleavage event occurred between amino acids Asn-1631 and Ser-1632, a dipeptide known to be susceptible to succinimide-based degradation[122,123].

6.4 The C-terminal domain of TraI exhibits a novel fold

The structured 1476-1629 region of the TraI-CT is composed of an α-Domain containing two α-helices at the N-terminus, a proline-rich loop, and a compact α/β-Domain at the C-terminus (**Figure 6.6**). The α/β-Domain contains two three-stranded β-sheets, a small β-hairpin loop and two helices. Antiparallel β-sheet 1 (β-strands 1-3) contains a β-hairpin loop inserted between β-strands 2 and 3. In β-sheet 2, β-strands 4 and 5 are antiparallel, while the insertion of α-helix 3 allows β-strand 6 to run parallel to β-strand 5.

TraI-CT crystallized as a domain-swapped dimer in the asymmetric unit. The N-terminal α-Domain of one monomer docks into the core domain of the second monomer (**Figure 6.7A**). β-sheet 1 forms a binding surface for helix 1 of the second monomer in the asymmetric unit, while the second β-sheet (β-strands 4-6) packs against helix 3. The α- and α/β-Domains are connected by a proline-rich loop, which is rigid relative to the remainder of the protein. The mean thermal displacement parameter (*B*-factor) for all protein atoms is 36.0 $Å^2$ (**Table 6.1**), while prolines 1523, 1525 and 1530 in this loop exhibit *B*-factors of 26.1, 25.8 and 26.1 $Å^2$, respectively. A crystallographic 2-fold axis of symmetry in the $C222_1$ space group further generates an intimately associated tetramer formed by two domain-swapped dimers (**Figure 6.7B**).

Two potential conformations of 1476-1629 region of the TraI-CT monomer can be hypothesized: an extended arrangement, in which one monomer is simply removed from the dimer (**Figure 6.7C**; see also **Fig. 6.6**), and an α + α/β globular arrangement, in which a single domain is generated by pairing the α -Domain of one monomer with the α/β-Domain of its domain-swapping partner (**Figure 6.7D**). Both the extended and the globular conformations were examined in DALI searches of the Protein Data Bank[124] for structurally-similar folds. Significant homology to TraI-CT was not detected in protein structure

similarity searches. Using DALI, two hits were observed for the extended form of TraI-CT, and three for the globular form. For both the extended and the globular conformations, the best match (with a DALI Z-score of 2.5 and 3.1 respectively) was a portion of *Pseudomonas putida* 2-oxoisovalerate dehydrogenase (PDB ID 1qs0). While this molecule contains a similar degree of α and β character, we discerned no structural similarity with the TraI-CT. Indeed, up to eleven separate regions of structural similarity between the TraI-CT and the oxidoreductase were observed; the aligned segments were 3 to 13 residues in length with root-mean-square deviations of 2.8 to 3.9 Å between Cα positions. After the oxidoreductase, the next best DALI matches had Z-scores of 2.2 and 1.7. Using secondary structure matching server (SSM)[125] and either the extended or globular forms of TraI-CT as a search model produced only three hits, all of which exhibited insignificant statistical values. These observations suggest that the 1476-1629 region of the F TraI C-terminal domain exhibits a novel fold.

6.5 DLS confirms the monomeric C-terminal structure

To investigate the polydispersity and oligomerization state of TraI constructs, dynamic light scattering experiments were employed using a Wyatt DAWN EOS light scattering instrument interfaced to an Amersham Biosciences Akta FPLC with Superdex S200 size exclusion column, a Wyatt Optilab refractometer, and Wyatt dynamic light scattering module. Constructs of TraI residues1476-1630 and 1476-1756 were expressed and purified using the protocol outlined in section 6.4. The samples were dialyzed into dynamic light scattering (DLS) buffer (10 mM potassium phosphate pH 7.1 and 150 mM NaCl). Data

were analyzed using the ASTRA software from Wyatt Technology. Molecular weights were calculated using the following equation:

$$\frac{K^* c}{R(\Theta)} = \frac{1}{MP(\Theta)} + 2A_2 c \qquad \textbf{Equation 6.1}$$

where K* is an optical parameter, c is the sample concentration, M is molecular weight, R(Θ) is the excess intensity of scattered light at DAWN angle Θ, P(Θ) is a function describing the angular dependence on scattered light, and A_2 is the second virial coefficient.

Both the complete C-terminus (1476-1756) and a construct containing the ordered region of the crystal structure (1476-1630) eluted as single peaks on the Superdex 200 column. Coupling size exclusion chromatography with dynamic light scattering (SEC-DLS) and refractometry facilitated the calculation of sample molecular weights. Both C-terminal constructs (1476-1756, 1476-1630) exist as homogeneous monomers with less than 3.5% difference between calculated and theoretical molecular weights (**Table 6.4**). It was noted that the theoretical molecular weight for the structured region of the TraI-CT (1476-1630), 16.7 kD, was smaller than the measured molecular weights from the SEC-DLS experiment (17.1–17.6 kD; **Table 6.4**). This observation suggests that the TraI-CT monomer may be in equilibrium between the extended and globular states (see **Figures 6.7C** and **6.7D**, respectively). Limited proteolysis using either trypsin or chymotrypsin generated a 3 kD product from the TraI-CT (data not shown). We hypothesize that this 3 kD fragment is the 28-residue α-domain, which is connected to the TraI-CT α/β-domain by the proline-rich loop composed of a sequence susceptible to cleavage by trypsin and chymotrypsin. This result further supports the conclusion that the TraI-CT monomer may shift between the closed and open states, the latter of which providing a substrate more amenable to proteolysis. The "far" C-terminal region (1630-1756), which was not present in our crystal structure, eluted in the

void volume (MW > 600 kDa) but does not precipitate in the standard aqueous buffers employed (data not shown), suggesting a soluble aggregate.

6.6 Transfer activity of TraI C-terminus truncation mutants and examination of specific structural features

DNA transfer assays were conducted to test the importance of distinct regions of the F TraI C-terminus in conjugation (full methods and materials can be found in primary citation[113]). A panel of C-terminal truncation mutants was generated by introducing a stop codon at positions 1504, 1524, 1550, 1600, 1630, 1680, or 1720. DNA transfer efficiency (the number of transconjugation events per donor cell) was evaluated and normalized to wild-type levels (**Figure 6.8**). Previous work establishes that TraI is essential for CDT; when a plasmid with no TraI gene (ΔTraI) is utilized as a control, no conjugation is detected[20,23]. Deletion of residues 1504-1756 or 1524-1756 resulted in a complete loss of conjugative transfer (**Figure 6.8 A**). Elimination of residues C-terminal to positions 1550, 1600 or 1720 resulted in a 100-fold decrease in the transfer efficiency, while removal of residues C-terminal to 1630 or 1680 resulted in a 1,000-fold reduction in transfer efficiency. Full activity is observed only with the complete protein; removal of just 36 residues reduces conjugative DNA transfer (CDT) by more than 100-fold (N1720 column in **Figure 6.8A**).

The importance of specific TraI-CT structural features was next examined using conjugative DNA transfer assays. First, three prolines (1518, 1523 and 1525) within the proline-rich loop were mutated simultaneously to glycine (**Figure 6.9 A, B**). Prolines 1523 and 1525 are completely conserved, and proline 1518 is highly conserved in related TraI protein sequences (see **Figure 6.4**). Together, they may impart rigidity to the 1517-1525

loop (see above for *B*-factor analysis). Mutation of the three prolines dropped CDT 100,000-fold (**Figure 6.9A**). Second, residues 1517-1525, the entire proline-rich loop, were deleted. Elimination of the 1517-1525 loop reduced CDT 10,000-fold (**Figure 6.9A**). Thus, the conformationally-restricted proline residues within this loop are essential for CDT.

Third, the importance of contacts between α1 and β-sheet 1 (h1/s1) was examined (**Figure 6.9C**). Mutation of residues V1478, E1482 and F1485 to alanine on α1, coupled with mutation of G1540 to glutamic acid and I1541 to alanine on β2 (h1/s1), resulted in only a 10-fold decrease in transfer efficiency relative to wild type TraI (**Figure 6.9A**). These mutations were designed to disrupt the interaction between α1 and β-sheet 2. Although we did not establish that this interaction was successfully disrupted (by measuring changes in protein stability, for example), the relatively moderate 10-fold decrease in transfer efficiency lead us to conclude that the α1/β-sheet 1 interaction is not essential to CDT.

Fourth, contacts between α3 and β-strands 4 and 6 of β-sheet 2 (h3/s2) were examined. In contrast to the α1/β-sheet 1 interaction, the mutation of residues L1574, Q1575 and V1603 (**Figure 6.9D**) simultaneously to alanine resulted in a >200-fold decrease in transfer efficiency (**Figure 6.9A**). Mutation of L1574 and V1603 to alanine is expected to disrupt the hydrophobic interactions between helix 3 and sheet 2. The Q1575 side chain nitrogen on β-strand 4 forms polar contacts with two main-chain oxygen atoms in the loop that packs helix 3 against β-strand 6, and also forms a polar contact with one main-chain oxygen on helix 3. When alanine replaces Q1575, these polar contacts would be eliminated. Thus, the globular nature of the α/β-domain of the TraI C-terminus plays an important role in conjugative DNA transfer. Taken together, these data functionally annotate the crystal

structure of the TraI C-terminal domain, and they establish that the α/β-domain and the rigid 1517-1525 loop are important for conjugative DNA transfer.

6.7 TraI C-terminal domain binds ssDNA

Because TraI contains domains that perform site-specific DNA nicking and highly processive DNA unwinding activities, the ability of constructs of the TraI C-terminal domain to bind to DNA was examined using fluorescence anisotropy. (Full materials and methods can be found in the corresponding publication[113]. The full TraI C-terminal domain (1476-1756) did not bind to a double stranded stretch of DNA (data not shown), in accordance with previously published data[112]. The C-terminal domain binds to a 34 nucleotide single-stranded DNA (ssDNA) oligo with K_ds of 2.9 μM and 7.7 μM in 75 mM and 150 mM NaCl, respectively (**Figure 6.10, blue**). A protein construct containing only the far C-terminal residues (1630-1756) formed a soluble aggregate in solution and was therefore not suitable for anisotropy studies. Similarly, the 1476-1629 fragment without the far C-terminal residues aggregated at 75 mM NaCl. At 150 mM NaCl, the 1476-1629 fragment exhibited poor binding to ssDNA (K_d>22.6 μM) (**Figure 6.10, black**). (A K_d value greater than 22.6 μM indicates a dissociation constant higher than the highest protein concentration tested.)

The proline loop deletion (Δloop) and Pro-to-Gly mutant (pmut) forms of the TraI C-terminal domain (1476-1756) were also examined. The Δloop form of TraI 1476-1756 binds ssDNA poorly, with K_ds of >17.1 μM and >15.9 μM at 75 mM and 150 mM NaCl, respectively (**Figure 6.10, red**). Similarly, mutating prolines 1518, 1523 and 1525 to glycine within this loop region also produced a form of TraI 1476-1756 that binds ssDNA poorly, with K_ds of >13.9 μM and >18.2 μM at 75 mM and 150 mM NaCl, respectively (**Figure**

6.10, green). As a negative control, we found that bovine serum albumin did not bind to ssDNA at either 75 mM or 150 mM NaCl (data not shown). These results demonstrate that the TraI C-terminal domain binds ssDNA, and indicate that both the 1517-1525 loop and the 1630-1756 region are essential for this activity. Thus, the presence and the relative structural rigidity of the 1517-1525 loop are required for ssDNA binding.

6.8 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 6. Figures taken directly from our publication have the citation in the figure title[113].

Figure 6.10 Binding of ssDNA by the TraI C-terminus measured by fluorescence anisotropy

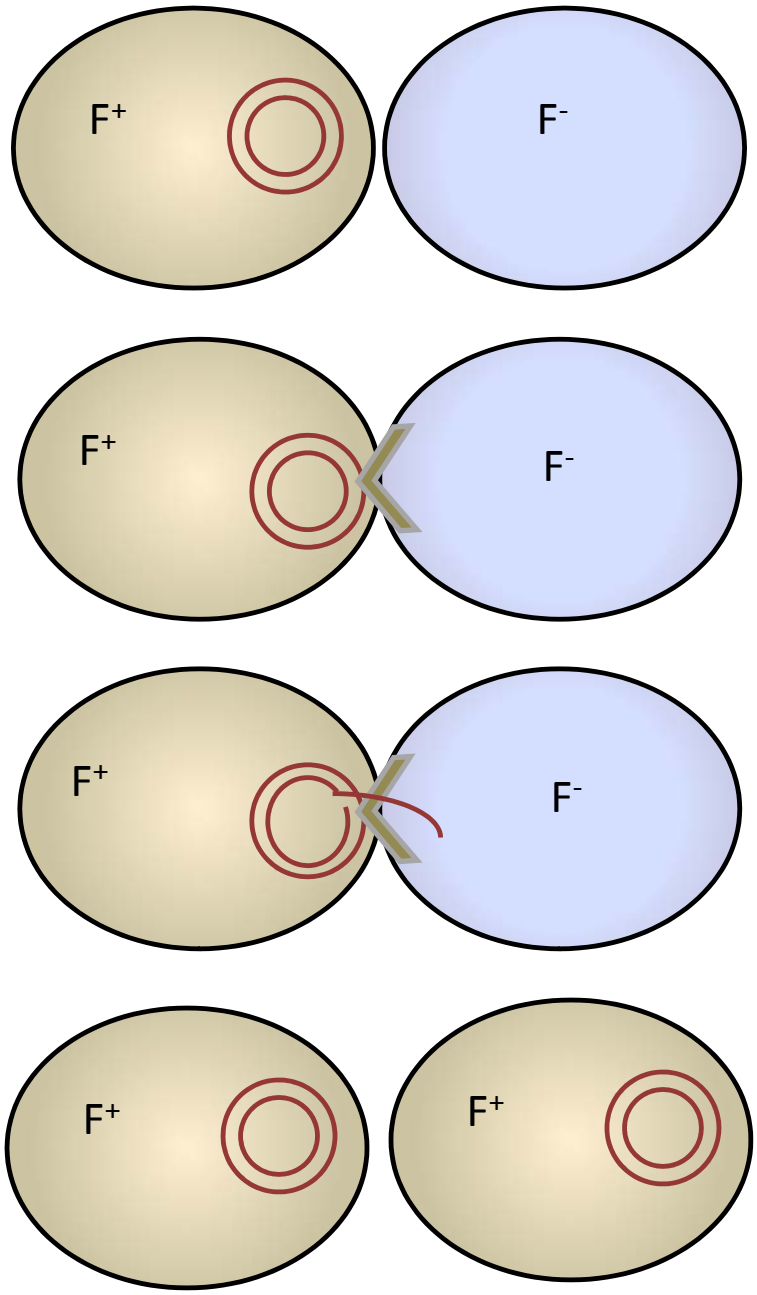


Figure 6.1 Simple model of conjugative DNA transfer used in the F plasmid system

An F^+ cell (brown) contacts an F^- cell (blue) through the type four secretion system (TFSS) (less than symbol). The relaxosome (not shown) assembles on the F plasmid (red circle) and nicks the DNA, unwinds the DNA and couples with the TFSS to transfer the strand to the F^- cell. The DNA is then replicated and an intact F-plasmid is now present in the new cell making it an F^+ cell.

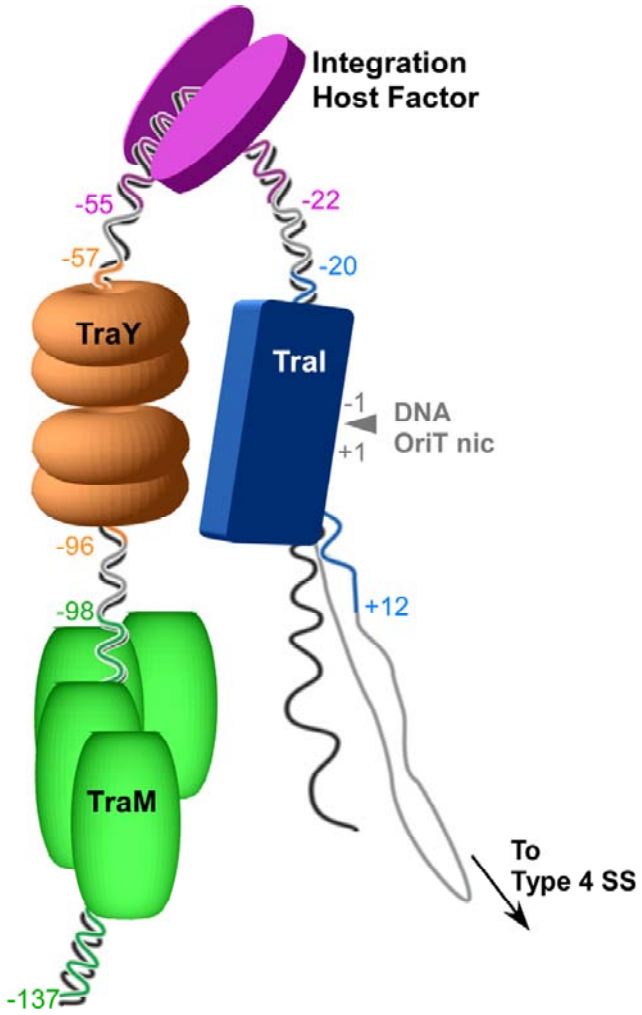


Figure 6.2 Schematic of the F-plasmid relaxosome[113]

The F plasmid-encoded proteins TraM (green), TraY (orange), TraI (blue) and the *E. coli* host encoded integration host factor (purple) bind the F plasmid DNA in a site- and sequence-specific manner. The nucleotides most proximal sites to the *nic* site (IHFA, SbyA, and SbmC) bound by each protein are indicated by color and numbered relative to the origin of transfer *nic* site (*oriT nic*). The region of ssDNA to be transferred to the recipient bacterial cell via a type IV secretion system (Type 4 SS) is shown.

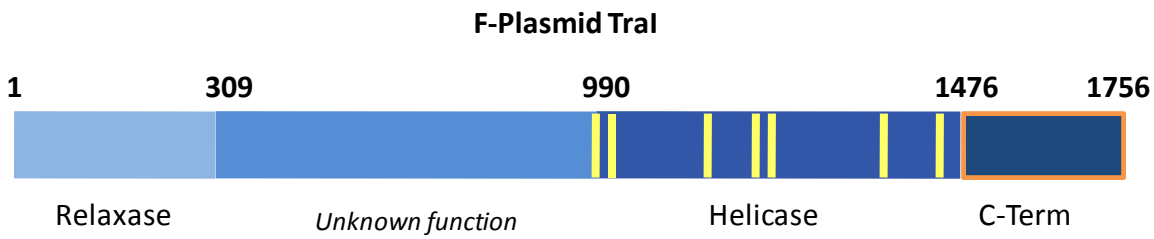


Figure 6.3 Domain structure of TraI[113]

The numbers represent amino acid positions. The N-terminal 300 residues compose the relaxase domain, while residues 990-1476 compose the helicase domain. The yellow bars represent canonical helical motifs. The 309-990 region has an unknown function and the 1476-1756 region makes up the C-terminal domain.

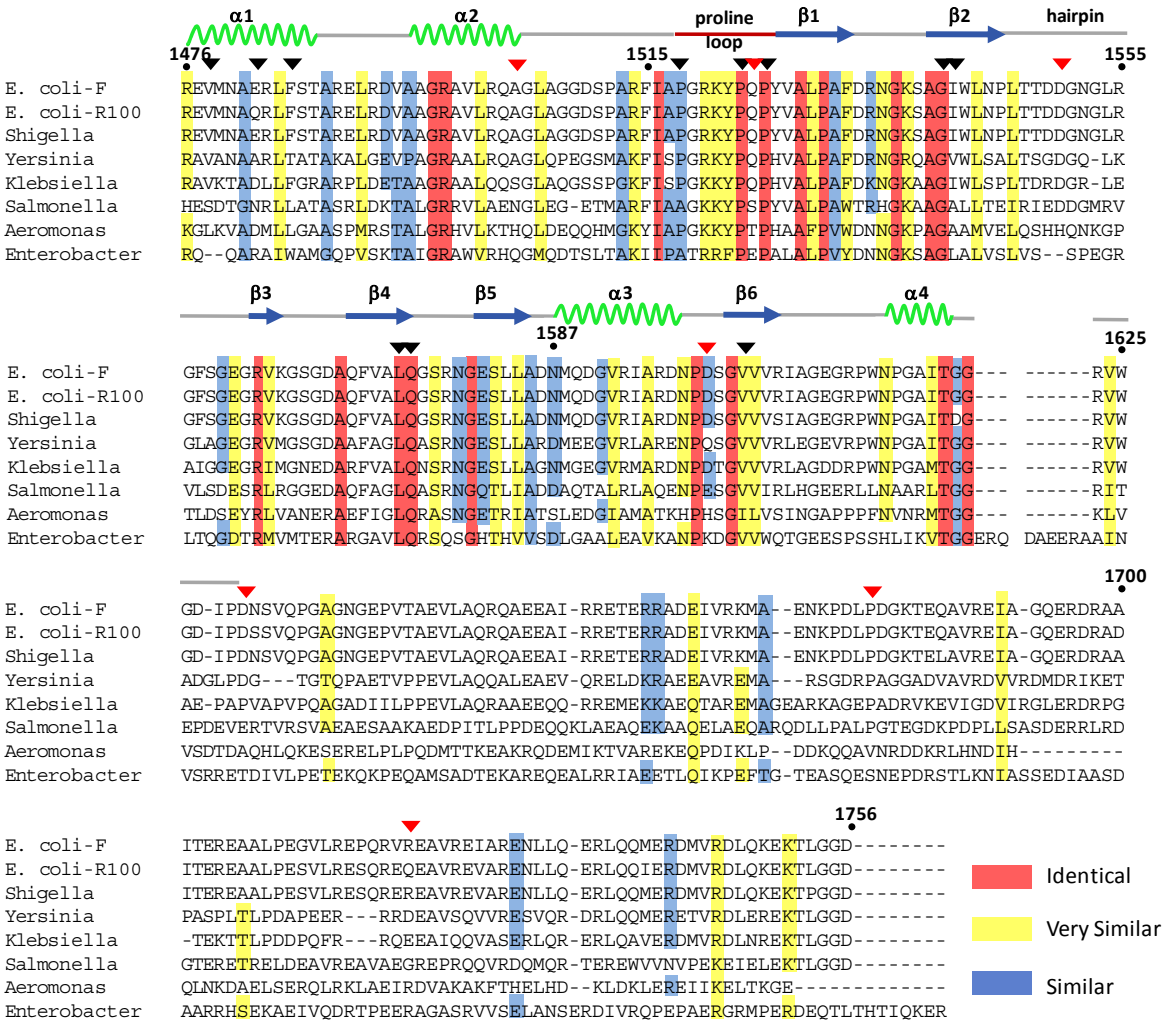


Figure 6.4 Sequence alignment of *E. coli* F-plasmid TraI C-terminus with TraI orthologs

Sequence alignment showing similar (blue), highly similar (yellow) and identical (red) residues. The sequence of the *E. coli* F plasmid TraI is compared to TraI orthologs from the *E. coli* plasmid R-100, as well as plasmids from *S. sonnei* (Ss046), *Y. pestis* (MT), *K. pneumoniae* (pMGH78587), *S. typhi* (pED208), *A. salmonicida* A449 (p5), and *Enterobacter* sp.638. The positions of deletion-generating stop codons described in Figure 5 are indicated with red triangles, and the positions of mutations described in Figure 6 are indicated with black triangles.

Table 6.1 Original statistics for the TraI C-terminal structure

Data collection				
X-ray source	APS SER-CAT BM-22			
Space group		$C222_1$		$P2_1$
Unit Cell[a] *a*, *b*, *c* (Å); α, β, γ (°)	40.8, 139.8, 126.8; 90,90,90			40.8,126.5,72.8; 90,106.2,90
	Peak	Inflection	Remote	Peak
Wavelength (Å)	0.97835	0.97873	0.97126	0.97835
Resolution (Å) (highest shell)	50.0-2.10 (2.18-2.10)	50.0-2.40 (2.48-2.40)	50.0-2.50 (2.59-2.50)	50.0-2.10 (2.18-2.10)
R_{sym}[b]	8.2 (28.6)	9.6 (42.5)	9.6 (41.9)	9.2 (37.7)
$I/\sigma I$	17.4 (2.8)	22.9 (3.7)	13.9 (2.3)	14 (2.6)
Completeness (%)	87.6 (60.5)	98.8 (92.9)	96.6 (81.3)	95.3 (80.4)
Redundancy	3.5 (2.9)	3.7 (3.2)	6.8 (6.0)	3.7 (3.3)
Phasing				
Mean Figure of Merit				
Sharp-Centric	0.301			
Sharp-Acentric	0.184			
Solomon & DM	0.804			
Refinement				
Resolution (Å)	50-2.10			
No. reflections	39344			
R_{work}[c]	0.247			
R_{free}[d]	0.283			
Molecules per asymmetric unit (AU)[a]	4			
No. of amino acids/AU	614			
No. of waters/AU	393			
No. of Ammonium sulfates/AU	6			
B-factors				
Protein	37.4			
Ammonium Sulfate	82.0			
Water	42.4			
R.m.s deviations				
Bond lengths (Å)	0.007			
Bond angles (°)	1.33			
Ramachandran (%)				
Favored, allowed, disallowed	95.31, 3.56, 0.83			

[a]Space group assignment changed during the course of refinement; see Experimental Procedures.

[b]$R_{sym}=\Sigma|I-I_{mean}|/\Sigma I$ where I is the observed intensity and I_{mean} is the average intensity of several symmetry related observations.

[c]$R_{work}=\Sigma|F_o-F_c|/\Sigma F_o$ where F_o and F_c are the observed and calculated structure factors, respectively.

[d]R_{free}=calculated as above for 5% of data not used in any step of refinement.

Table 6.2 Redundancy, I/σ and completeness for data scaled to 2.1 Å or 2.4 Å resolution

	$C222_1$ scaled to 2.1 Å		$C222_1$ scaled to 2.4 Å	
Statistic	**Shell (Å)**	**Value**	**Shell (Å)**	**Value**
Average redundancy per shell	2.26-2.18	5.6	2.59-2.49	6.3
	2.18-2.10	5.2	2.49-2.40	6.1
	All	6.6	All	6.9
Percent of total observed reflections with I/σ of <2	2.26-2.18	53.1	2.59-2.49	45.8
	2.18-2.10	57.2	2.49-2.40	46.8
	All	31.1	All	23.4
Percent of total observed reflections with I/σ of <10	2.26-2.18	87.0	2.59-2.49	78.3
	2.18-2.10	89.5	2.49-2.40	81.7
	All	59.8	All	50.4
% Completeness (% reflections observed)	2.26-2.18	71.6	2.59-2.49	91.5
	2.18-2.10	63.6	2.49-2.40	89.4
	All	89.3	All	97.5

Table 6.3 Final data collection, phasing, and refinement statistics[113]

Data collection

X-ray source	APS SER-CAT BM-22		
Space Group	$C222_1$		
Unit cell: a,b,c (Å); α, β, γ (°)	40.8, 139.8, 126.5; 90, 90, 90		
	Peak	Inflection	Remote
Wavelength (Å)	0.97835	0.97873	0.97126
Resolution (Å) (highest shell)	50.0-2.39 (2.49-2.39)	50.0-2.40 (2.49-2.40)	50.0-2.50 (2.59-2.50)
R_{sym}	8.8 (27.6)	9.6 (42.5)	9.6 (41.9)
I/σ	28.6 (5.4)	22.9 (3.7)	13.9 (2.3)
Completeness (%)	97.5 (89.4)	98.8 (92.9)	96.6 (81.3)
Redundancy	6.9 (6.10)	3.7 (3.2)	6.8 (6.0)

Phasing

Mean Figure of Merit	
Sharp-Centric	0.301
Sharp-Acentric	0.184
Solomon & DM	0.804

Refinement

Resolution (Å)	50.0-2.4
No. reflections	24509
R_{work}	0.220
R_{free}	0.263
Molecules per asymmetric unit (AU)	2
No. of amino acids per AU	307
No. of waters per AU	130
No. of sulfates per AU	10
B-factors	
Protein	36.0
Sulfates	90.6
Water	38.1
R.M.S. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.4
Ramachandran (%)	
Favored	96.4
Outliers	0

$R_{sym}=\Sigma|I-I_{mean}|/\Sigma I$ where I is the observed intensity and I_{mean} is the average intensity of several symmetry related observations.

$R_{work}=\Sigma|F_o-F_c|/\Sigma F_o$ where F_o and F_c are the observed and calculated structure factors, respectively.

R_{free}=calculated as above for 5% of data not used in any step of refinement.

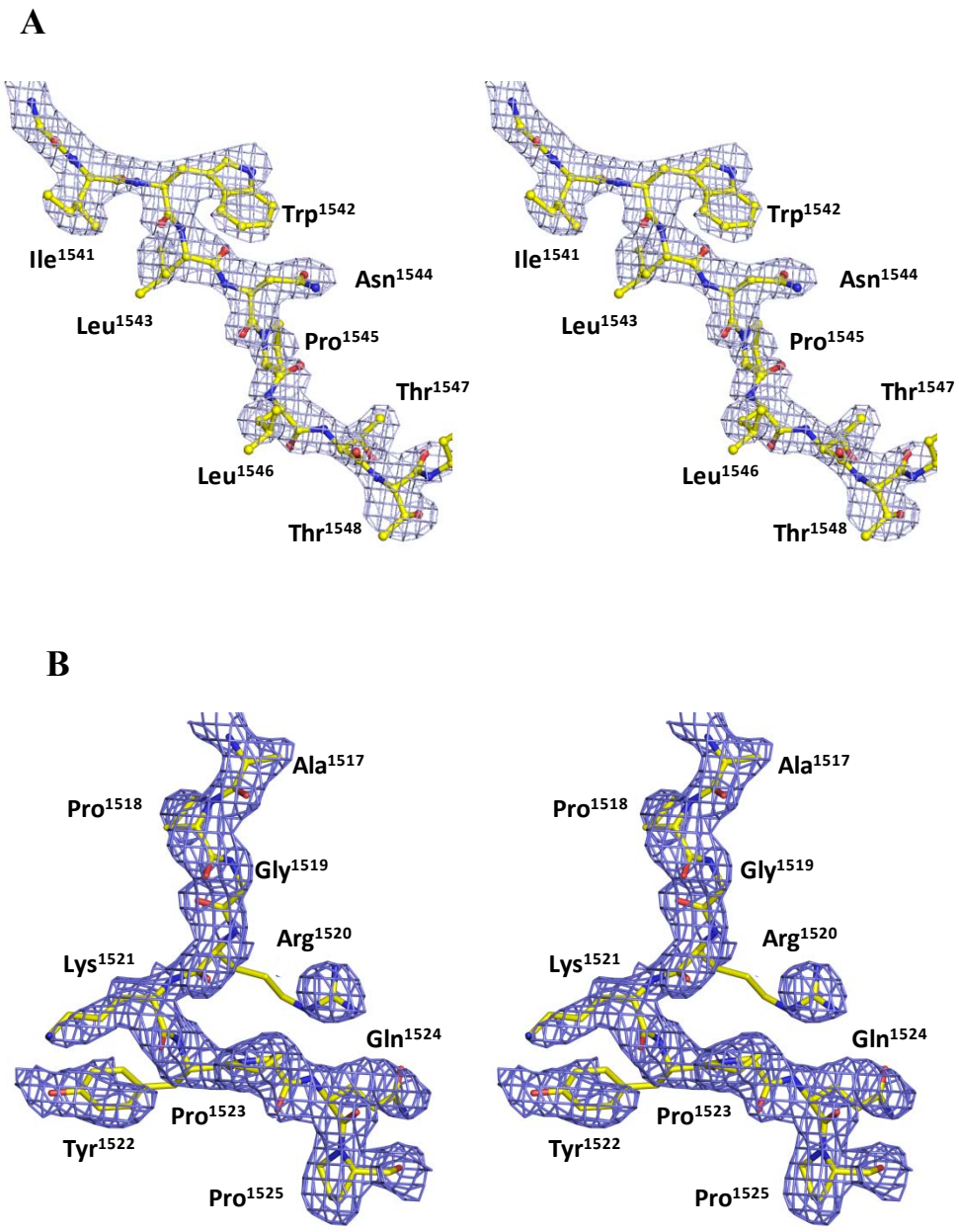


Figure 6.5 Two portions of the final model with the experimental electron density from MAD phasing[113]

Stereoview of two portions of the original 2.4 Å resolution solvent-flattened experimental electron density map after SHARP, SOLOMON and DM (contoured at 1.5 σ with the final refined protein model. Residues isoleucine 1541 through threonine 1548 are illustrated in (**A**), while (**B**) shows the proline rich loop between residues alanine 1517 and proline 1525.

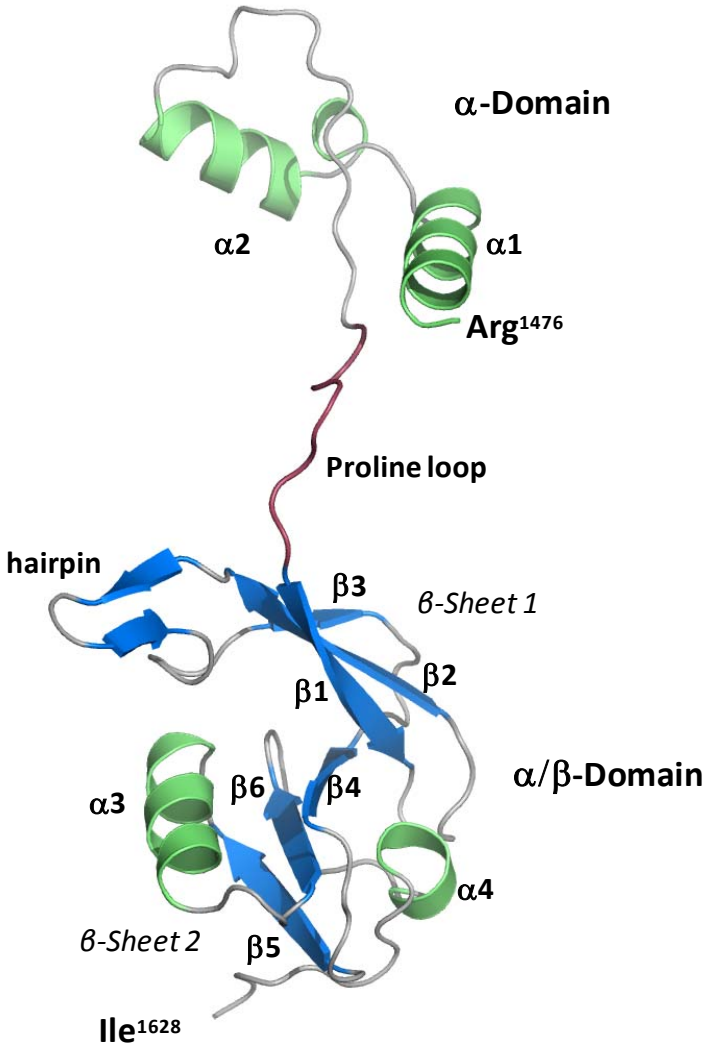


Figure 6.6 TraI C-terminal structure[113]

Crystal structure of a monomer of the F plasmid TraI-CT. Secondary structure elements are indicated in green (helices), blue (β-strands), grey (loops) and red (proline-rich loop).

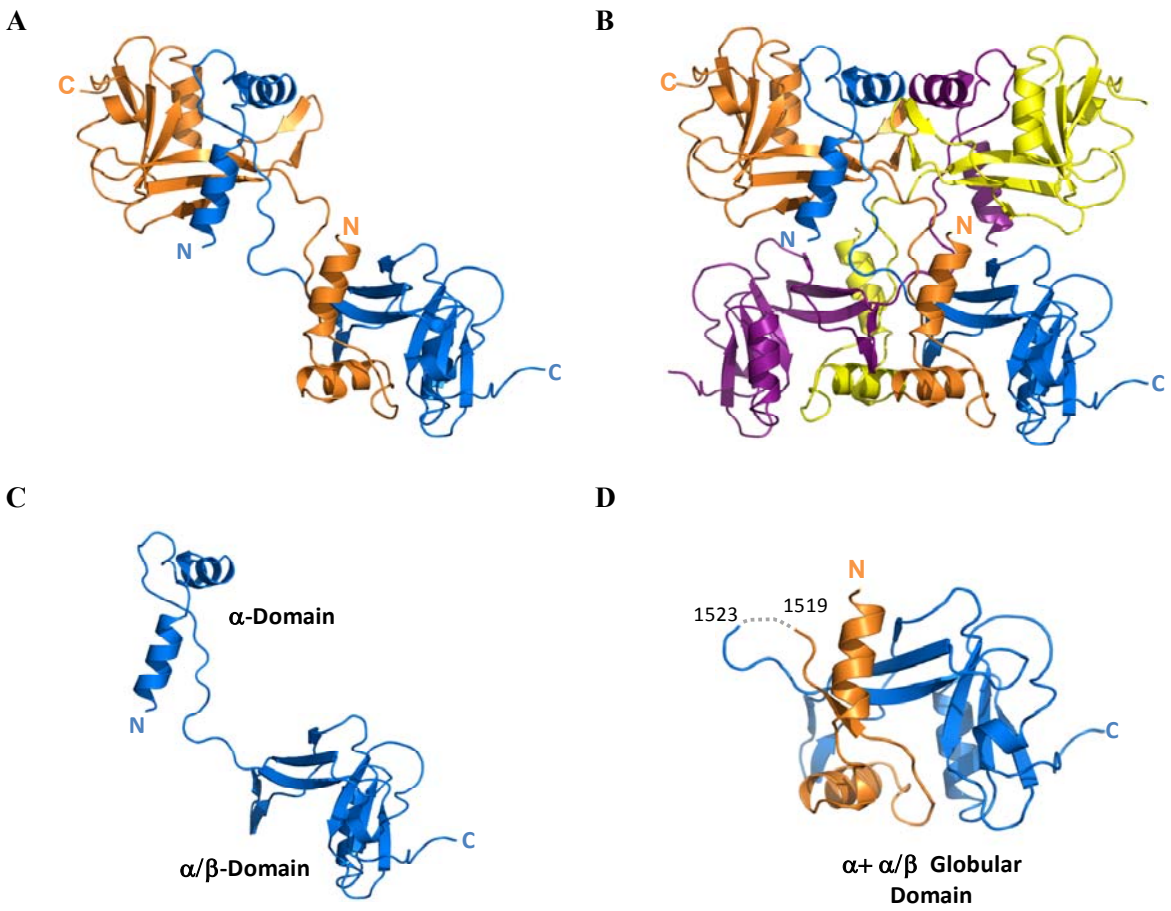


Figure 6.7 Dimer, tetramer and monomer representations of the TraI C-terminal structure[113]

(**A**) TraI-CT forms a domain-swapped dimer in the asymmetric unit. Each monomer is oriented with the N-terminus (α-Domain) of one monomer contacting the C-terminus (α/β-Domain) of the other monomer. (**B**) A domain-swapped tetramer is also generated between symmetry-related dimers. Each monomer is oriented with the N-terminus towards the core of the tetramer and the C-terminus at the edge (orange interacts with blue, and purple with yellow). Two monomeric forms of the TraI C-terminus can be modeled: extended (**C**) and globular (**D**).

Table 6.4 Size exclusion chromatography and dynamic light scattering[113]

Protein Construct	**TraI 1476-1756**	**TraI 1476-1630**
Theoretical MW (kDa)	31.180	16.690
Molar Mass Moments (kDa)		
Mn (% error)	32.31 (0.11%)	17.08 (0.5%)
Mw (% error)	32.32 (0.11%)	17.35 (0.5%)
Mz (% error)	32.33 (0.25%)	17.61 (0.5%)
Polydispersity		
Mw/Mn (% error)	1.000 (0.2%)	1.016 (0.7%)
Mz/Mn (% error)	1.001 (0.3%)	1.041 (91.1%)

*Weight average molar mass defined as $Mw=\Sigma(c_i \cdot M_i)/ \Sigma c_i$, Number average molar mass defined as $Mn= \Sigma c_i / \Sigma(c_i/M_i)$, Z-average molar mass defined as $Mz = \Sigma(c_i \cdot M_i)^2/ \Sigma(c_i \cdot M_i)$.
Polydispersity of the sample equals one only when the sample has homogenous molecular mass (i.e. one oligomeric state and independent of averaging method).

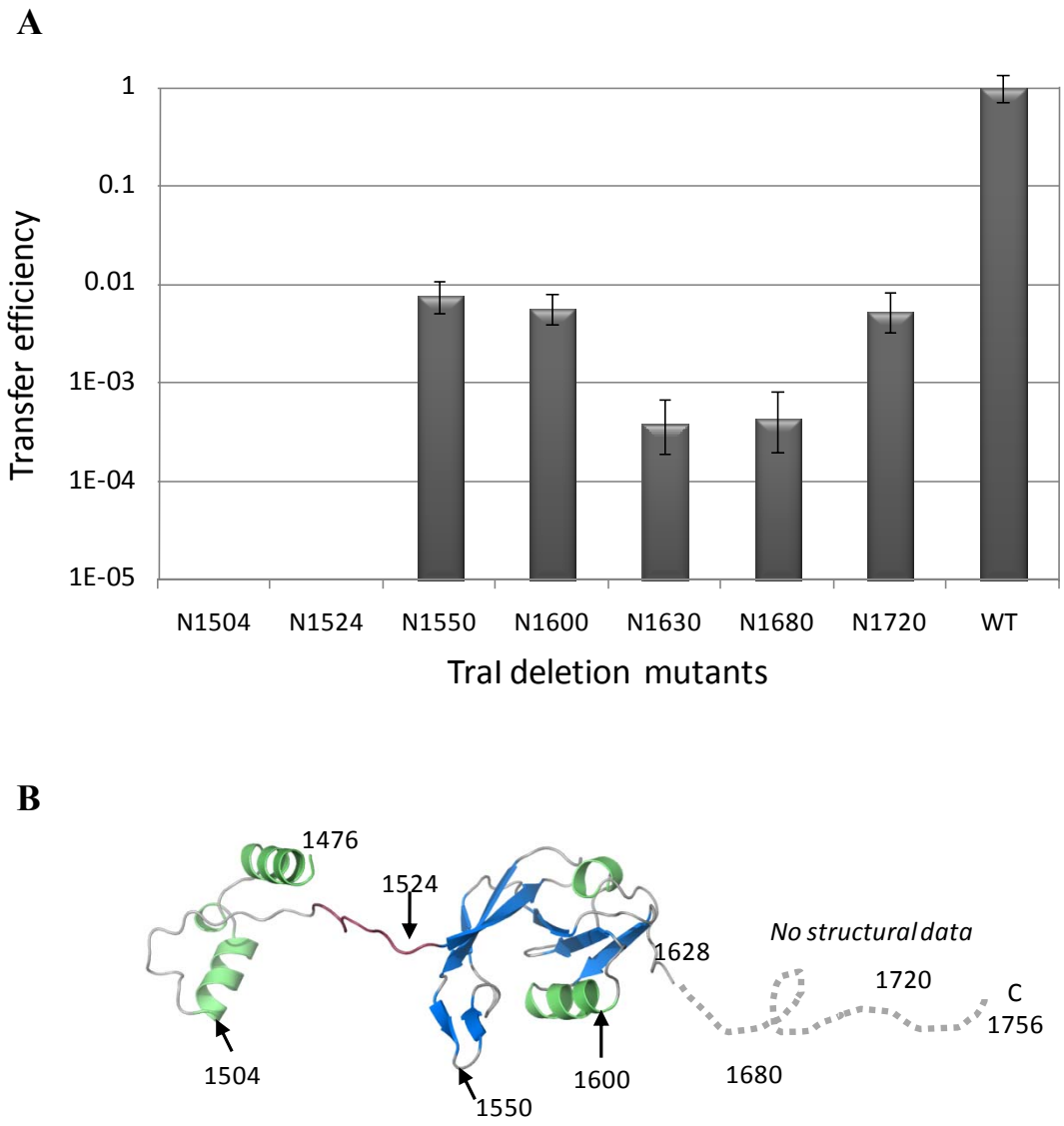


Figure 6.8 Transfer efficiency of TraI C-terminus deletion mutants[113]

(**A**) Conjugative DNA transfer efficiencies of TraI deletion mutants. Constructs include residues N-terminal to the indicated residue; for example, N1550 includes residues 1-1550. (**B**) The TraI-CT structure annotated with the stop codon sites used for this truncation analysis.

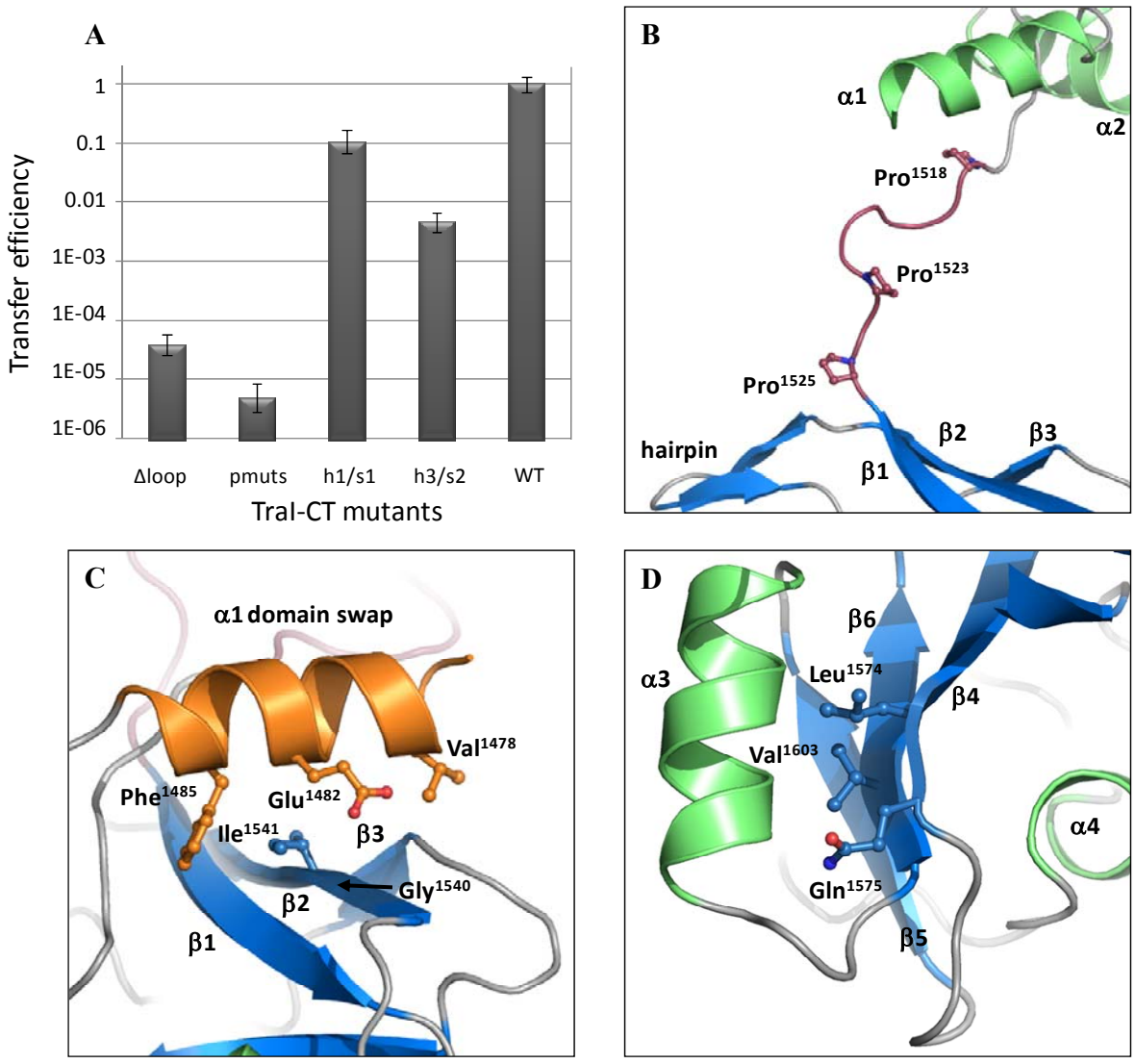


Figure 6.9 Plasmid transfer efficiency using specifically designed mutant TraI proteins to examine contacts within the C-terminal structure[113]

(**A**) Conjugative DNA transfer efficiencies of bacterial strains containing specifically designed TraI-CT mutations. (**B**) "Pmuts" indicates the mutation of prolines 1518, 1523 and 1525 simultaneously to glycine, while "Δloop" indicates the removal of the entire proline-rich loop. (**C**) The helix 1/sheet1 (h1/s1) variant is the mutation of V1478, E1482 and F1485 all to alanine on helix 1 and the mutation of I1541 to alanine and G1540 to glutamic acid on strand 2. (Helix 1 is shown here in orange to indicate the domain swapped interaction.) (**D**) The helix 3/sheet2 (h3/s2) mutants replace L1574, Q1575 and V1603 all with alanine.

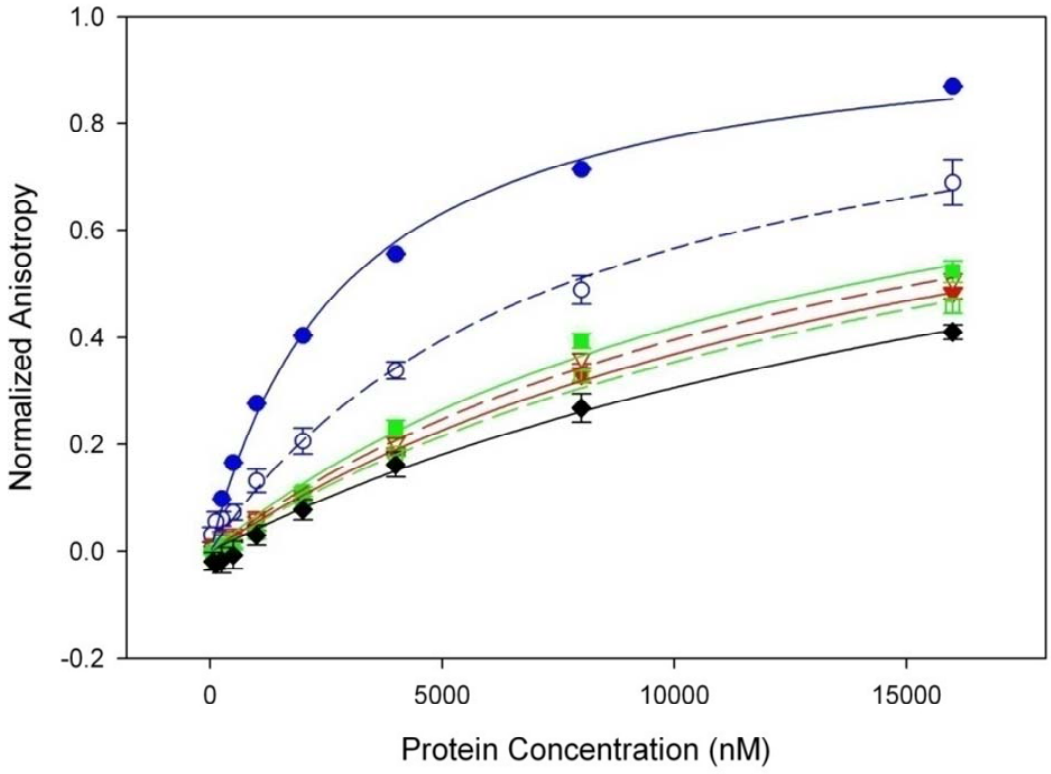


Figure 6.10 Binding of ssDNA by the TraI C-terminus measured by fluorescence anisotropy[113]

TraI 1476-1756 at 75 and 150 mM NaCl is indicated by solid and dashed blue lines (K_d =2.9 μM and K_d =7.7 μM) respectively. TraI 1476-1756 with a deletion of the proline rich loop at 75 and 150 mM NaCl is indicated by solid and dashed red lines (K_d >17.1 μM and K_d >15.9 μM) respectively. TraI 1476-1756 with mutations of prolines 1518, 1523 and 1525 to glycine at 75 and 150 mM NaCl is indicated by solid and dashed green lines (K_d >13.9 μM and K_d >18.2 μM) respectively, while the binding of 1476-1630 at 150 mM NaCl is indicated in black (K_d >22.6μM).

Chapter 7: Other (unpublished) projects

7.1 Initial characterization of TraI putative GTPase

Investigation of the *E. coli* F-plasmid TraI protein has been a central focus for the Redinbo laboratory since 2004, when Scott Lujan started collaborating with Dr. Matson in the UNC-CH Biology Department. Scott crystallized the N-terminal relaxase domain of TraI and demonstrated its role in bacterial conjugative DNA transfer through inhibiting plasmid tranfer[109]. Chapter 6 introduces this protein and discusses the structure and function of the C-terminal portion of TraI[113]. The central region of the protein had not yet been assigned any functional role, but we identified a putative GTPase in the central region of TraI, encompassing residues 309-590. I started working on this part of the protein as a training project with my second undergraduate student, Sung Taek Kim. The project goal was to determine if this central region is in fact an enzyme with GTPase activity and to determine which residues are the most important for GTP binding. The following sections discuss our preliminary work.

7.1.1 Primary sequence analysis of TraI's putative GTPase

Examination of the primary sequence of TraI reveals amino acid sequences that are found in typical GTPases, including a Walker A box, Walker B box, and a nucleotide-specific box[126,127]. **Table 7.1** lists the canonical residues for each motif and the residues found in TraI. The lowercase x represents any amino acid and lowercase h denotes a hydrophobic residue. The Walker A box is responsible for binding the γ-phosphate of the

NTP, while the Walker B motif is used for coordinating the metal ion used in the cleavage reaction[126-128]. The nucleotide-specific box is responsible for coordinating the specific nitrogenous base. TraI has a nucleotide-specific box that reflects a GTP binding motif, not an ATP binding motif.

To visualize the structure of these types of motifs, I performed a PHYRE search against TraI to obtain structures that are predicted to be similar to this central portion of TraI[48,129]. Protein FFH from *T. aquaticus* is the closest structural homologue predicted for TraI. **Figure 7.1** shows the Walker A motif of FFH, which does have the canonical residues GxxxxGKT. This structure shows a metal bound in the active site and also shows a GDP molecule. In this Walker A motif, as in others, the G residues are for loop flexibility, K is to orient the γ-phosphate, and T is to help orient the metal ion. TraI's putative Walker A box does have multiple G residues that could provide flexibility, a positively charged R that could replace the K, and a polar negatively charged E that could replace the T. As seen in **Figure 7.2**, FFH has a GTP specific box with a TKxD motif and has a GDP molecule bound to this site. The K and D residues are important for making hydrogen bonds to the ribose and the guanine. TraI has both K and D residues in the nucleotide-specific region for binding the guanine base. **Table 7.1** shows that there are two potential Walker B boxes present in TraI. We hypothesized that TraI does contain a GTPase region in the 309-590 region and planned assays and mutagenesis to address this hypothesis.

7.1.2 Assay development for TraI GTPase

Originally TraI's GTPase activity was investigated using a coupled assay that monitored the production of the fluorescent compound resorufin (**Figure 7.3**). This assay was adopted from an experiment that was used to measure the helicase activity of the Rec

family[130]. The assay involves several enzymes, including pyruvate kinase, pyruvate oxidase and horseradish peroxidase. The NDP generated from the GTPase is utilized by pyruvate kinase to convert PEP to pyruvate. In the presence of oxygen, pyruvate oxidase converts pyruvate to acetylphosphate while generating H_2O_2. Then, horseradish peroxidase utilizes the peroxide to convert amplex red into resorufin. However, resorufin is very insoluble in water, and thus I was not able to get a reliable standard curve. To avoid the complications with resorufin, I found an alternative assay for probing GTPase activity.

The next generation of the assay, adapted from Gosselin *et al*. utilized one less coupled enzyme and monitored the oxidation of NADH to NAD^+ rather than the production of resorufin[131]. The first step is the same: NDP conversion to NTP by pryuvate kinase, which drives PEP conversion to pyruvate. Then, lactate dehydrogenase converts pyruvate to lactate in the presence of NADH; NADH is converted to NAD^+. **Figure 7.4** shows the schematic of this assay. The readout of this assay was initially monitored by NADH fluorescence emission at 460 nm. However, the dynamic range of the fluorescent readout was very limiting, so NADH absorbance at 350 nm was utilized as the indication of GTPase activity.

Figure 7.5 shows the standard curve relating absorbance with concentration of NADH. The decrease in absorbance caused by NADH conversion to NAD^+ is directly proportional to the amount of NTP converted to NDP. Several control experiments were performed to make sure that the coupled enzymes (LDH and PK) were not limiting factors in the experiment. An excess of these enzymes was always utilized. Also, control reactions were performed to make sure that a change in absorbance would only be seen when all components of the reaction were present. As seen in **Figure 7.6**, control reactions were monitored where individual components were separately removed from the reaction. Each of

these control reactions did not have significant change in absorbance signal over time. The full length wild-type is shown on this graph as a positive control reaction. Also, the reaction with no NADH had very low absorbance at 350 nm, indicating that no other assay components absorbed at this wavelength.

7.1.3 Mutation design and protein expression

As described previously, the TraI GTPase Walker boxes are not typical, so we were interested in identifying the specific residues that are important for activity. An extensive number of constructs were cloned and expressed through mine and Sung Kim's efforts. **Table 7.2** shows all of the constructs and mutations that were designed for investigating this putative GTPase. Construct 309-590 represents only the GTPase, 309-958 includes the GTPase and a region of unknown function, 309-1504 includes the canonical helicase and 1-1756 is the full length protein (see **Figure 6.3**). Each mutation was made using site-directed mutagenesis in the LIC-His plasmid and protein was expressed in BL21 cells. Proteins were purified using nickel affinity and size exclusion chromatography. Purification buffers are similar to those described in the methods section in chapter 2. Protein was concentrated to 80 μM and flash frozen in liquid nitrogen.

7.1.4 Initial results and redirection of project

The assay was performed in a 96 well block and read in the PheraStar (BMG). Each well contained a master mix of 0.7 mM PEP, 0.6 mM NADH, 0.1% BSA, 10 mM MgAc, 25 mM Tris Acetate, pH 7.5 and 4 mM NADH and water. Then, the appropriate amount of GTP or ATP was added to have a range of 0.04-40 μM GTP or ATP. Data collection began immediately after the addition of 2 μM protein. Background absorbance was subtracted and then the NADH absorbance was converted into NADH concentration through the standard

curve equation (**Figure 7.5**) and plotted versus time. The initial linear portion of the line was selected as the initial velocity for the reaction and could then be plotted versus substrate concentration to find the Km and Vmax for each protein.

Figure 7.7 illustrates the first experimental evidence for the existence of the TraI GTPase. This data indicates that residues G445, G449, Q450 and R451 are important for maintaining wild-type level activity. Further work done by undergraduate Sung Kim has shown that although there is a basal level of GTPase activity in the 309-590 region, but the canonical helicase from residues 900-1500 is the main ATPase and does not require the GTPase portion for full activity. Thus, the extensive mutational studies of the GTPase are no longer a priority. In retrospect, I have learned that developing a sound biochemical assay and fully characterizing the wild-type constructs should be done before any mutational analysis.

While Sung and I worked on this part of TraI, Dr. Mike Miley was working on the helicase portion of the proteins. As this work became more intertwined through DNA binding assays and activity assays, Mike continued this project while I solved the TraI C-terminal structure and finished the Symplekin HEAT domain project. Thus, this GTPase characterization has been combined with other lab member's efforts. Currently, Yuan Cheng is working towards solving the crystal structure of the full length TraI protein and also finding the ideal piece of DNA for crystallization.

7.2 Progress towards structural characterization of human Eppin

For a few months while I was exploring project options, I worked on a project pertaining to male contraception, a field that I find especially interesting. Dr. Michael O'Rand, of the Cell and Developmental Biology Department at UNC-CH, is currently

leading a team of researchers to investigate structure and function of human spermatozoa proteins. The goal of the collaborative work was to produce recombinant Eppin protein in sufficient quantity and purity for crystallization and then to solve the structure of the full length or a segment of the Eppin protein. I worked in collaboration with their laboratory for a short time before committing to the Symplekin and TraI projects.

7.2.1 Introduction

Immunization of male monkeys by production of Eppin antibodies was en effective, reversible method of contraception: seven out of nine male monkeys developed high anti-Eppin titer and became infertile, upon ceasing immunization, five of the seven became fertilet[132]. This was the main result that made Eppin an interesting structural target. Eppin (**ep**ididymal **p**rotease **in**hibitor) coats the human ejaculate spermatozoa and has been shown to bind to semenogelin (Sg), which is the main protein constituent of seminal fluid. Binding of Eppin to Sg protects the fluid from microbes and also from degradation by proteases[133,134]. These findings make the structural and functional study of Eppin and Sg imperative to understand the molecular level mechanism of the male contraception.

7.2.2 Cloning, expression and purification of human Eppin

The DNA for Eppin and semenogelin was received from the O'Rand laboratory. The genes were then cloned into the LIC-MBP plasmid for further characterization. **Table 7.3** lists the primers used in cloning the full length and structural domains of these proteins. The 134 residue Eppin protein is predicted to contain an N-terminal (residues 1-72) WAP domain similar to the elastase inhibitor, elafin (PDB 2REL) and a C-terminal (73-128) domain homologous to Kunitz-type tryspin inhibitor (PDB 1KTHa)[133]. Both of these domains are held in place by multiple cysteine bonds, as can be seen in **Figure 7.8**. The full length and

WAP domains were over-expressed in *E. coli* BL21 Origami cells to produce soluble proteins. Typically, *E. coli* cytoplasm is a reducing environment, so Origami cells were chosen because they have mutations in both thioredoxin reductase (*trxB*) and glutathione reductase (*gor*) genes to enhance disulfide bond formation in the cytoplasm. The proteins were purified using a Ni-affinity column, followed by TEV cleavage to remove the MBP tag, and then run again on the Ni-affinity column to purify away the His-MBP tag. The expression was verified with a western blot using Eppin antibodies; **Figure 7.9** shows the protein after TEV cleavage with amido black staining and western blotting. The primary antibody for the western blot was rabbit anti-eppin, and the secondary antibody is goat anti-rabbit. The near full length construct (18-133) in lanes 1 and 2 show the presences of bands at molecular weights 14, 28 and 56 kD, which likely corresponds to monomer, dimer and tetramer species. The Kunitz domain (75-133) in lane 3 was not present in either gel. The WAP domain (18-76) was present in bands at 7, 14 and 49 kD, representing possible monomer, dimer and heptamer species.

The complicating factor for moving forward with crystallization was the presence of these multiple oligomerization states for the full length protein and WAP domains. As seen in **Figure 7.8**, both domains are proposed to be held together by multiple disulfide bonds in the structures. The expression of the human protein in bacterial cells does not guarantee the correct formation of these disulfide bonds. Especially since both domains contain multiple cysteine residues, it is difficult to determine if the correct bonds are being formed within each domain. Experiments with reducing agents BME and DTT were able to reduce high molecular weight oligomers of Eppin 18-133 to dimer and monomer species, but the proteins were still a mixture of two dimer species and one monomer species as seen on an SDS-gel

(**Figure 7.10**). Also, chemical reduction may disrupt the necessary cysteine bonds present in both Kunitz and WAP domains. To avoid the oligomerization problem, the next course of action for this project was to develop a native purification for Eppin out of human ejaculate or to produce the protein in insect cells. However, at this time, the collaboration on the Symplekin project was gaining traction, so I was unable to complete the Eppin project. Hopefully another graduate student can continue this work.

7.3 Figures and Tables

Figures and tables are listed in the same order as they appear in the text of chapter 7.

Table 7.1 Canonical residues for NTPases and TraI residues indicating the presence of a GTPase

Motif	Canonical	TraI	TraI Residues
Walker A box	GxxxxGKT	GqGGaaGQRE	443-451
Walker B box	hhhhDxxG	TVIVDqG or VLITDsG	504-510 or 532-538
GTP specific box	TKxD	MKqD	476-479
ATP specific box	[N/T]SxQ	none	n/a

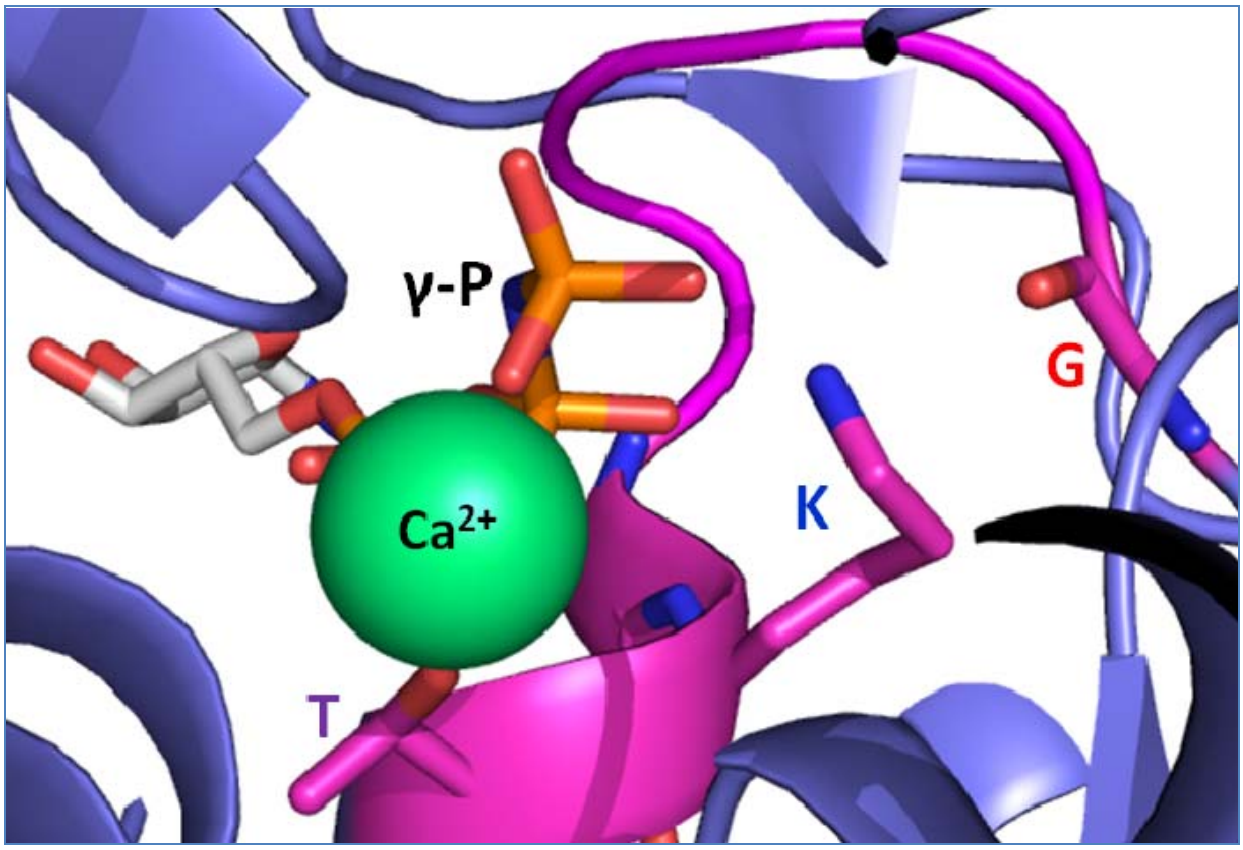


Figure 7.1 Walker A box of FFH (PDB 1ng1)

FFH is a GTPase from *Thermus aquaticus*. Here is a view of the Walker A box that is coordinating the GDP molecule. The γ-phosphate would be positioned by coordination with the threonine and lysine residues. A calcium metal is also coordinated in this region. FFH has the canonical walker A motif: GxxxxGKT.

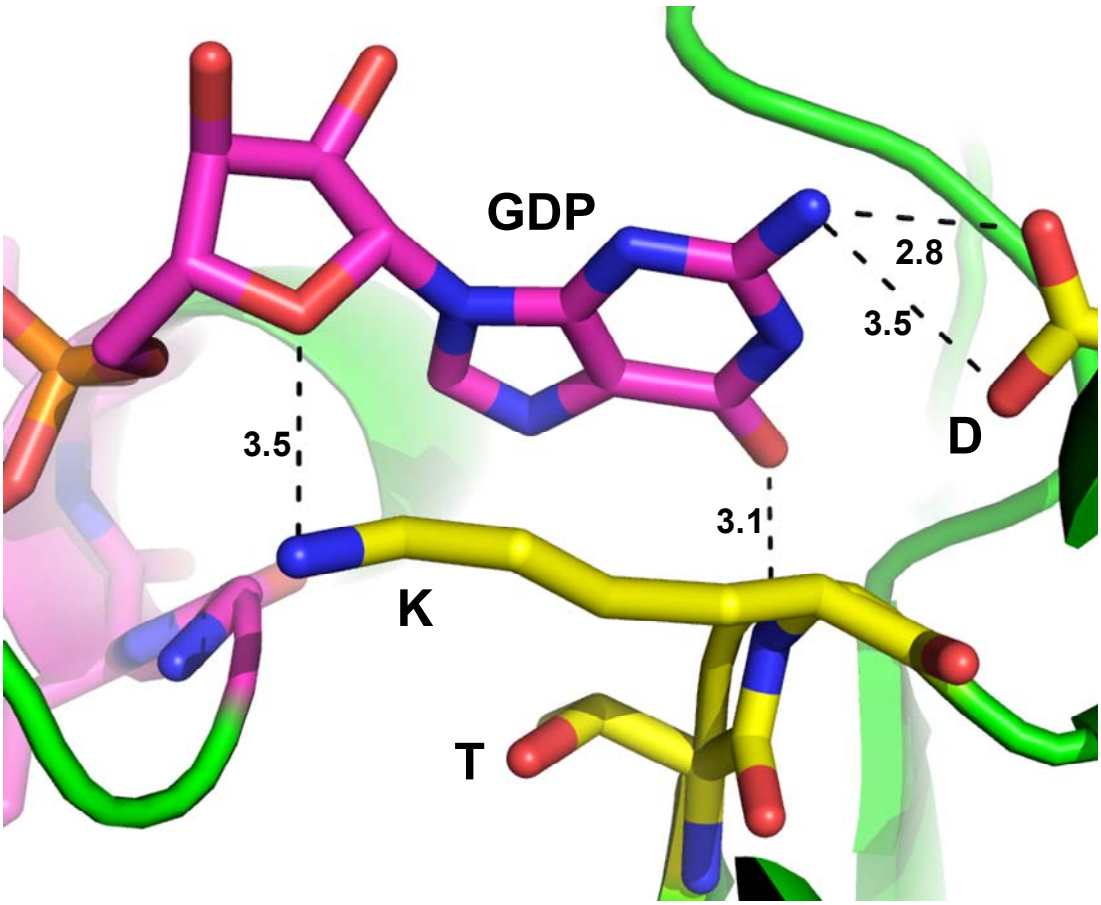


Figure 7.2 Nucleotide-specific box of FFH (PDB 1ng1)

Aspartic acid forms two hydrogen bonds with the NH_2 of guanine. The oxygen of guanine forms a hydrogen bond with the nitrogen backbone atom of the conserved lysine residue. The conserved lysine side chain nitrogen forms a hydrogen bond with the ribose sugar.

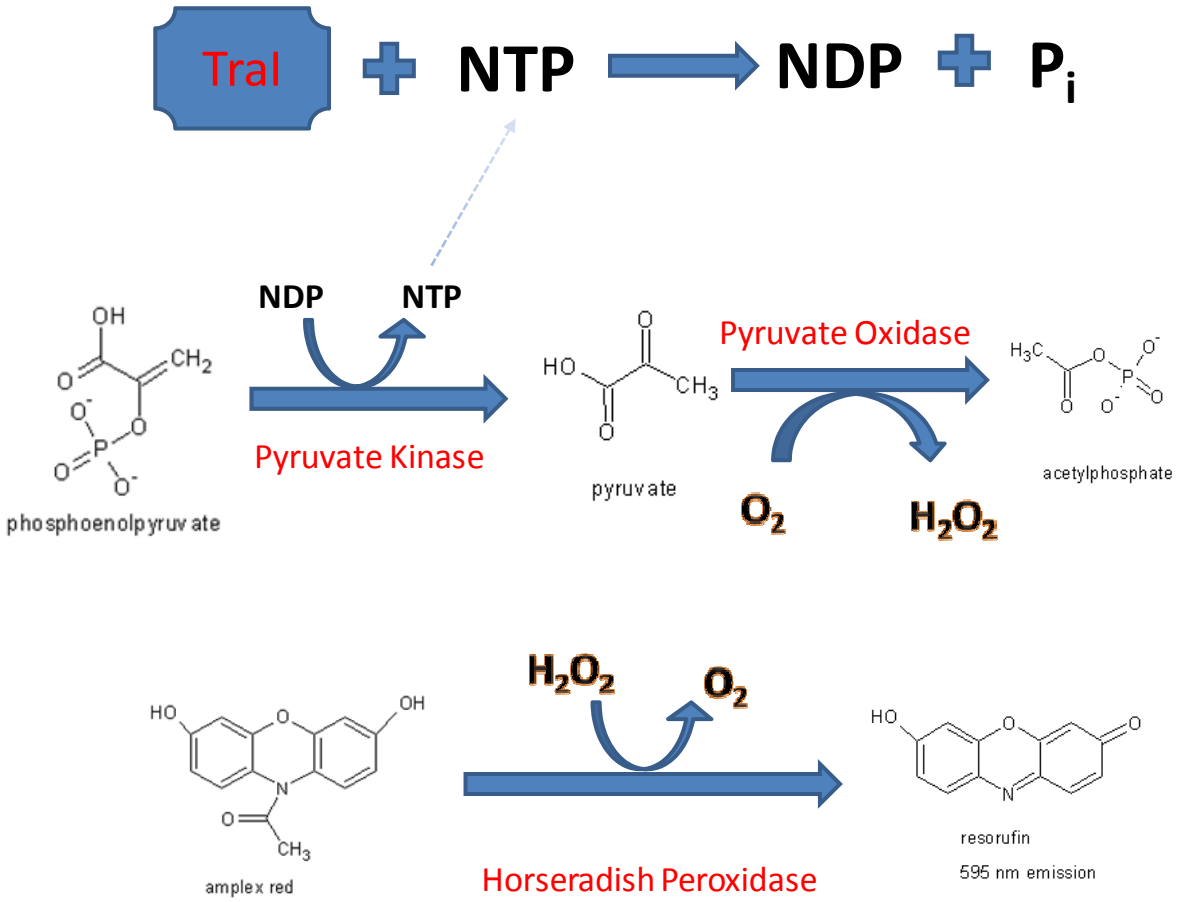


Figure 7.3 First generation NTPase assay utilizing resorufin

An NTPase will convert ATP to ADP and Pi. Then, the coupled reaction with pyruvate kinase, pyruvate oxidase and horseradish peroxidase will convert PEP through several intermediates including resorufin, which emits 595 nm light.

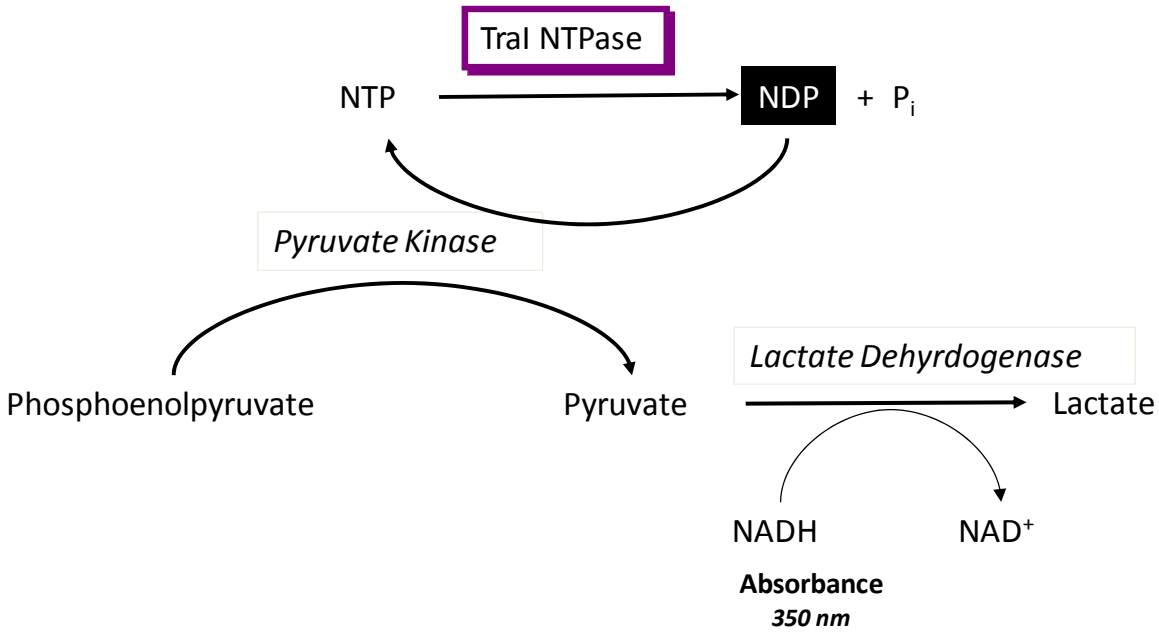


Figure 7.4 NTPase assay utilizing NADH absorbance

NTPase catalyzes the production of NDP from NTP. Pyruvate kinase converts NDP to NTP while converting PEP to pyruvate. Pyruvate is then converted to lactate in the presence of NADH. NADH is converted to NAD^+ during this final reaction. Monitoring the loss of absorbance at 350 nm indicates the NTPase activity of TraI.

Table 7.2 List of constructs and mutations to investigate the TraI GTPase

Construct	Mutation	Cloned	Grown	Purified	Done?
309-590	G443A	x	x	x	x
309-590	G445A	x	x	x	x
309-590	G446A				
309-590	G449A				
309-590	Q450A	x	x	x	x
309-590	R451A	x	x	x	x
309-590	M476A				
309-590	K477A	x	x	x	x
309-590	D479A	x	x	x	x
309-590	WT	x	x		
309-958	G443A	x	x		
309-958	G445A	x	x		
309-958	G446A				
309-958	G449A	x	x		
309-958	Q450A	x	x		
309-958	R451A				
309-958	M476A	x	x	x	x
309-958	K477A	x	x	x	x
309-958	D479A	x	x		
309-958	WT	x	x		
309-1504	G443A	x	x	x	x
309-1504	G445A	x	x	x	x
309-1504	G446A	x	x	x	x
309-1504	G449A	x	x	x	x
309-1504	Q450A	x	x	x	x
309-1504	R451A	x	x	x	x
309-1504	M476A	x	x	x	x
309-1504	K477A				
309-1504	D479A	x	x		
309-1504	WT	x	x		
1-1756	G443A				
1-1756	G445A				
1-1756	G446A				
1-1756	G449A				
1-1756	Q450A				
1-1756	R451A				
1-1756	M476A				
1-1756	K477A				
1-1756	D479A				
1-1756	WT	x	x		

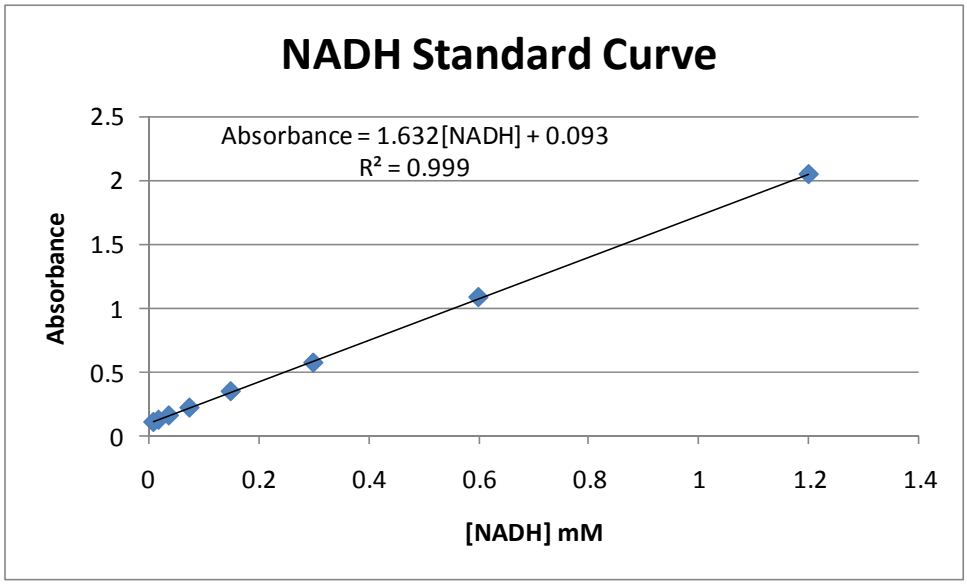


Figure 7.5 Standard curve of NADH

Standard curve of NADH concentration vs. absorbance. R^2 value of 0.999 indicates linearity. The equation can be rearranged so that any given absorbance can correlate to NADH concentration.

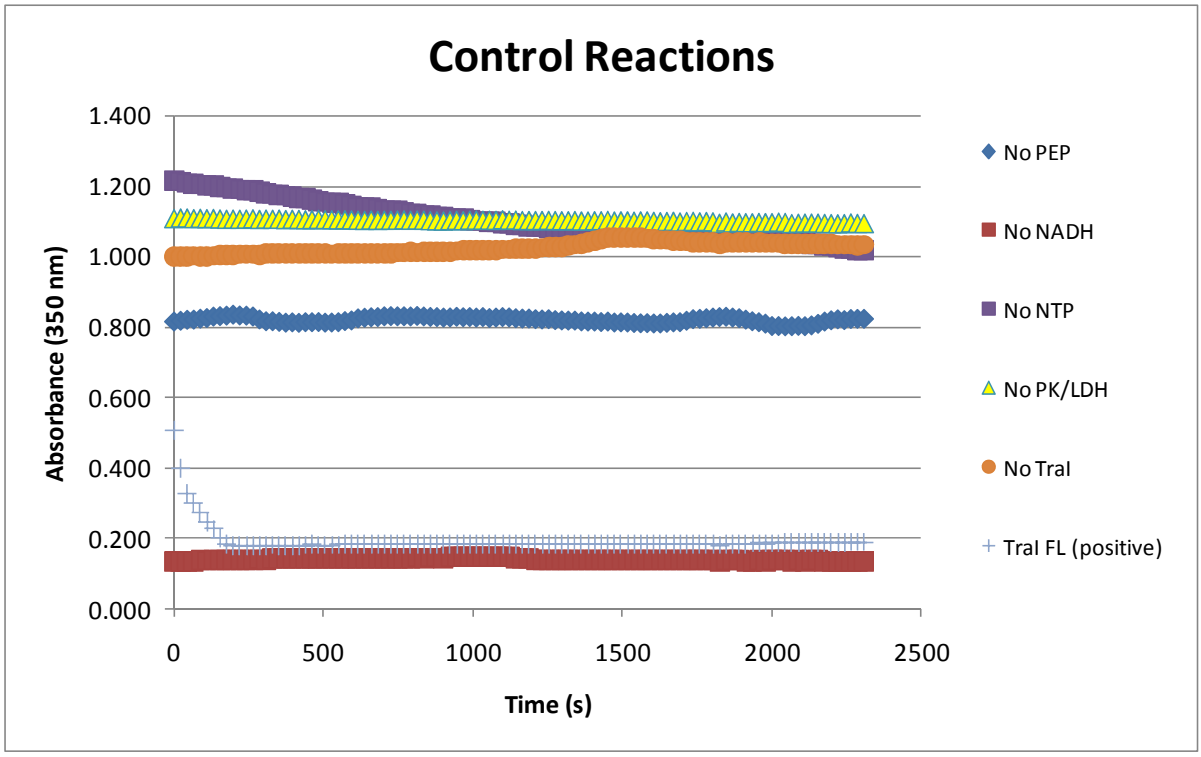


Figure 7.6 Control reactions for GTPase assay

Control reactions were performed to make sure that there was no change in absorbance signal in the absence of each required component. As expected, the reaction with no NADH has a very low absorbance. There is no significant change in absorbance value for reactions without PEP, TraI, PK/LDH or NTP. Full length TraI was added as a positive control to see the significant difference between the negative control reactions and the most active form of the wild-type enzyme.

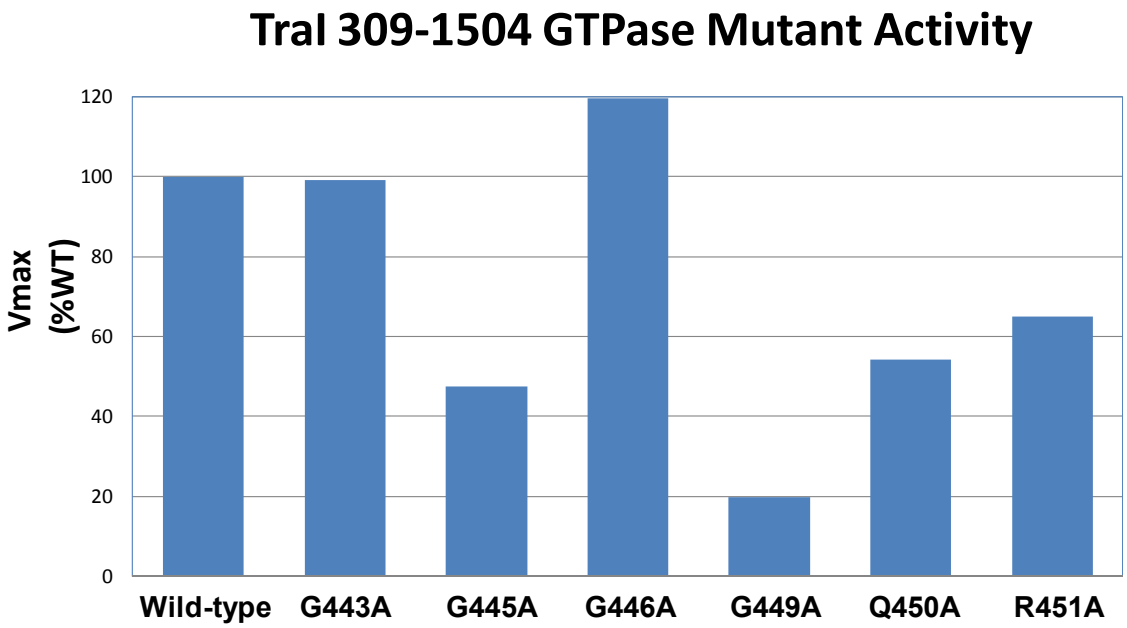


Figure 7.7 Initial GTPase activity of select mutants in the 309-1504 TraI construct

V_{max} (% of wild-type) of w TraI 309-1504 mutant constructs. This construct includes the putative GTPase and the canonical helicase regions. Assay conditions are 2 uM TraI protein, 0.03-4 M GTP substrate, 1mM Tris-acetate pH 7.5, and no DNA. The 309-1504 Wild Type Vmax = 9.18 nM/s, Km = 0.03 mM.

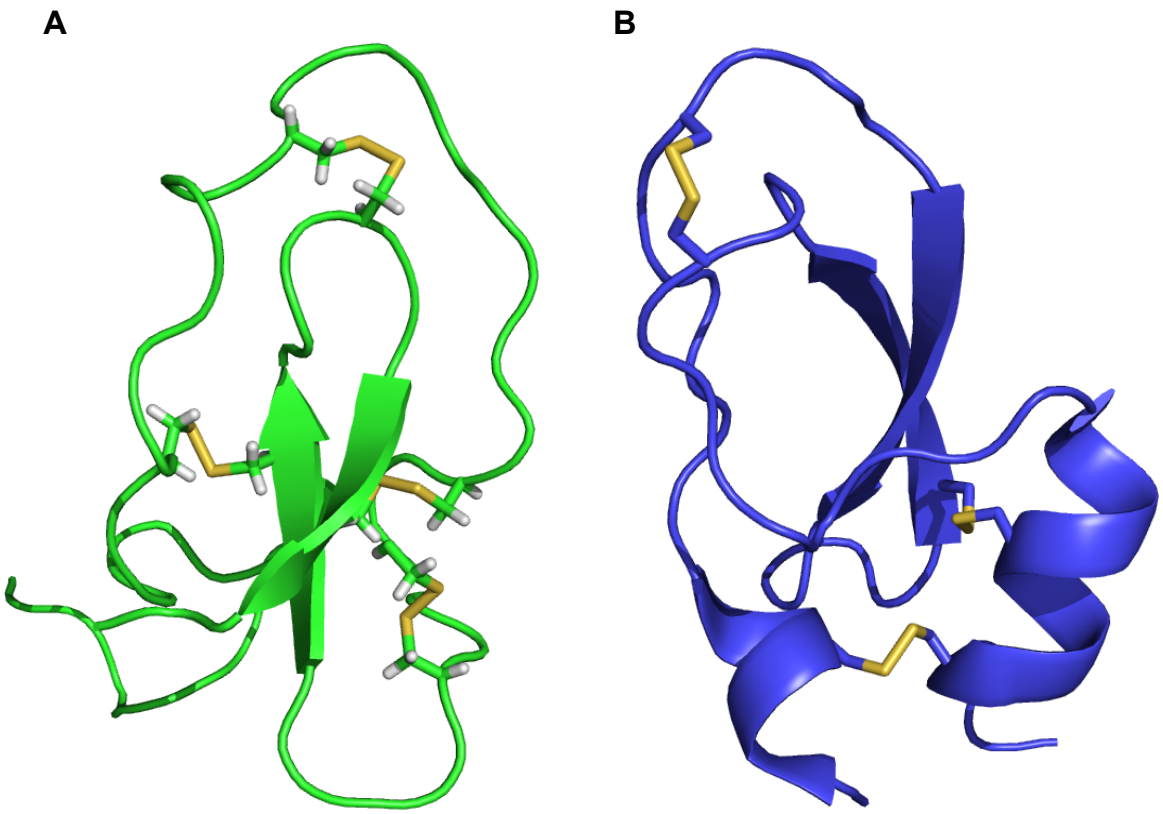


Figure 7.8 WAP and Kunitz domains of homologues proteins to Eppin

(A) WAP domain from elafin (PDB 2REL). The four cysteine bonds (yellow) are shown in stick representation. The N-terminal region of Eppin is predicted to be structurally similar to this WAP domain. (B) Kunitz domain of Kunitz-type domain C5 of collagen alpha 3 (VI chain. Three disulfide bonds (yellow) are shown in stick representation. The C-terminal region of Eppin is predicted to be structurally similar to this Kunitz domain.

Table 7.3 Primers for Eppin and Semenogelin for LIC cloning

Protein	Residue	F/R	Name	5'-3' sequence
Eppin	18	F	Ep18FLC	TAC TTC CAA TCC AAT GCG AAT GTC CAG GGA CCT GGT CTG
Eppin	75	F	Ep75FLC	TAC TTC CAA TCC AAT GCG GAT GTA TGC GAA ATG CCA AAA
Eppin	133	R	Ep133RLC	TTA TCC ACT TCC AAT GCG CTA GGG AAA GCG TTT ATT CTT GCA
Eppin	76	R	Ep76RLC	TTA TCC ACT TCC AAT GCG CTA TAC ATC TTG TTT GAG ATC TAA
Sg	26	F	Sg26FLC	TAC TTC CAA TCC AAT GCG GGT GGA TCA AAA GGC CGA TTA
Sg	164	F	Sg164FLC	TAC TTC CAA TCC AAT GCG AGG CTG TGG GTT CAT GGA CTA
Sg	282	F	Sg282FLC	TAC TTC CAA TCC AAT GCG AAG GCA AAT AAA ATA TCA TAC
Sg	265	R	Sg265RLC	TTA TCC ACT TCC AAT CGG CTA TGT AAA TAA TGG GTT TCG GTC
Sg	283	R	Sg283RLC	TTA TCC ACT TCC AAT CGG CTA TGC CTT TCG GCC ATG CTG TTG
Sg	401	R	Sg401RLC	TTA TCC ACT TCC AAT CGG CTA TTC ACC ATG CCA TGG CTC TTG

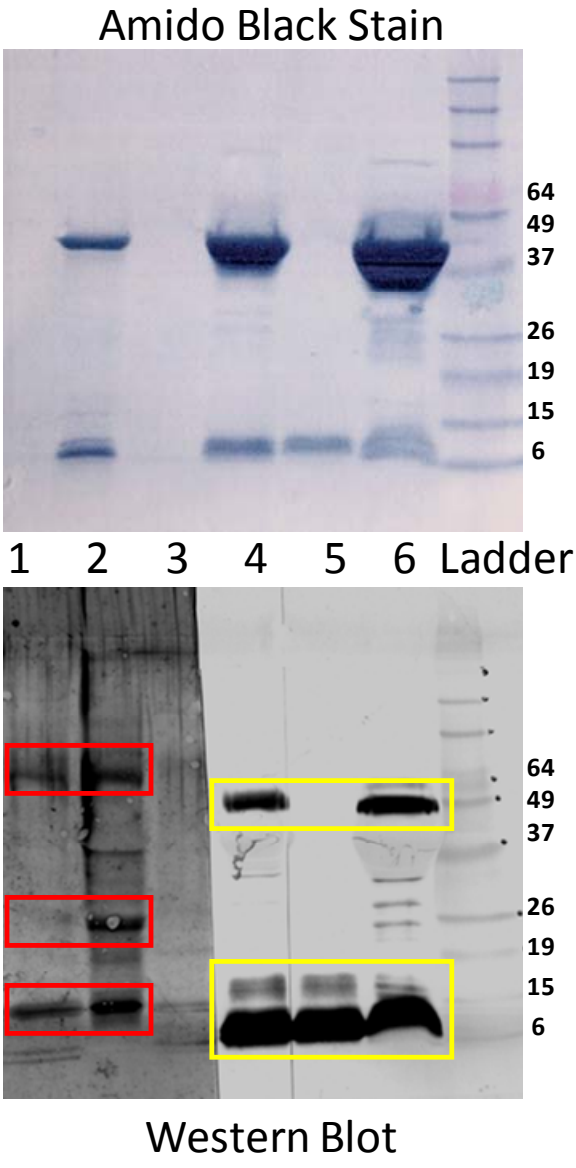


Figure 7.9 Expression of recombinant human Eppin in BL21 Origami cells

Amido black staining and western blot of human Eppin produced in BL21 Origami cells.
Lane 1: Eppin 18-133 (14 kD monomer) post TEV cleavage flow through.
Lane 2: Eppin 18-133 (14 kD monomer) post TEV cleavage elution.
Lane 3: Eppin 75-133 (7 kD monomer) post TEV cleavage flow through.
Lane 4: Eppin 18-76 (7 kD monomer) post TEV cleavage reaction (not purified).
Lane 5: Eppin 18-76 (7 kD monomer) post TEV cleavage flow through.
Lane 6: Eppin 18-76 (7 kD monomer) post TEV cleavage elution.
Western blot analysis shows lanes 1 and 2 contain bands at 56 kD (tetramer) 28 kD (dimer) and 14 kD (monomer). Lanes 4, 5, and 6 contain bands at 7 kD (monomer) and a series of bands between 7-14 kD, perhaps representing dimerization. Lanes 4 and 6 also have bands at 49 kD, which could represent a heptamer.

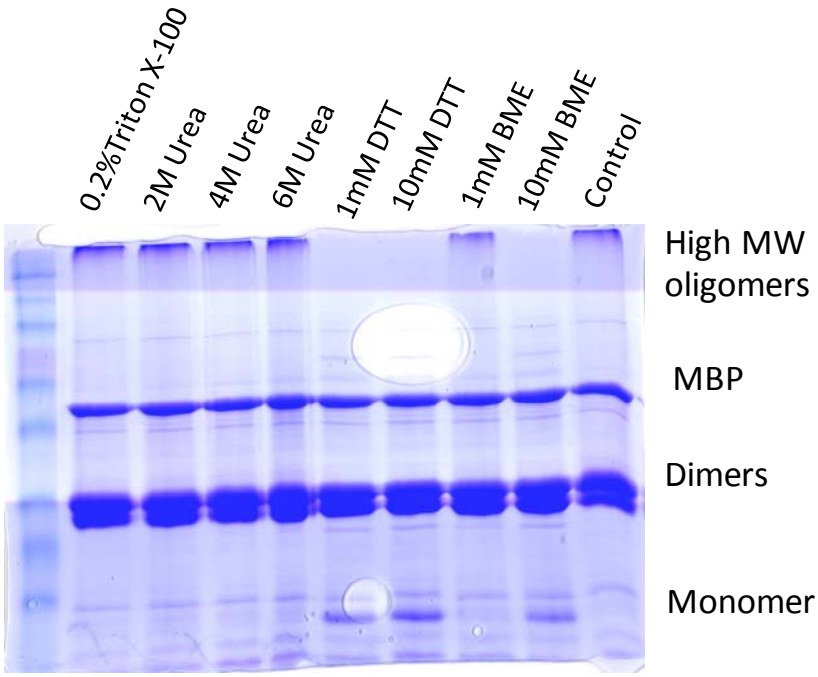


Figure 7.10 Eppin oligomerization is disrupted by reducing agents

Eppin 18-133 was subject to conditions to attempt to disrupt oligomerization. Eppin was resistant to Triton and Urea. DTT (1 or 10 mM) and 10 mM BME disrupted high molecular weight oligomerization and produced monomer sized protein. The two dimers were not reduced by either DTT or BME.

REFERENCES

1. Zhao, J., Hyman, L. & Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63, 405-445.

2. Mandel, C. R., Bai, Y. & Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci 65, 1099-1122.

3. Takagaki, Y. & Manley, J. L. (1998). Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell 2, 761-771.

4. Ford, L. P., Bagga, P. S. & Wilusz, J. (1997). The poly(A) tail inhibits the assembly of a 3'-to-5' exonuclease in an in vitro RNA stability system. Mol Cell Biol 17, 398-406.

5. Hunt, A. G., Xu, R., Addepalli, B., Rao, S., Forbes, K. P., Meeks, L. R., Xing, D., Mo, M., Zhao, H., Bandyopadhyay, A., Dampanaboina, L., Marion, A., Von Lanken, C. & Li, Q. Q. (2008). Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. BMC Genomics 9, 220.

6. Beelman, C. A. & Parker, R. (1995). Degradation of mRNA in eukaryotes. Cell 81, 179-183.

7. Proudfoot, N. J., Furger, A. & Dye, M. J. (2002). Integrating mRNA processing with transcription. Cell 108, 501-512.

8. Huang, Y. & Carmichael, G. G. (1996). Role of polyadenylation in nucleocytoplasmic transport of mRNA. Mol Cell Biol 16, 1534-1542.

9. Colgan, D. F. & Manley, J. L. (1997). Mechanism and regulation of mRNA polyadenylation. Genes Dev 11, 2755-2766.

10. Calvo, O. & Manley, J. L. (2003). Strange bedfellows: polyadenylation factors at the promoter. Genes Dev 17, 1321-1327.

11. Dominski, Z. & Marzluff, W. F. (2007). Formation of the 3' end of histone mRNA: getting closer to the end. Gene 396, 373-390.

12. Mandel, C. R., Bai, Y. & Tong, L. (2007). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci.

13. Sheets, M. D., Ogg, S. C. & Wickens, M. P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. Nucleic Acids Res 18, 5799-5805.

14. Wickens, M. & Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. Science 226, 1045-1051.

15. McLauchlan, J., Gaffney, D., Whitton, J. L. & Clements, J. B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Res 13, 1347-1368.

16. Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L. & Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature 444, 953-956.

17. Dettwiler, S., Aringhieri, C., Cardinale, S., Keller, W. & Barabino, S. M. (2004). Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. J Biol Chem 279, 35788-35797.

18. Ryan, K. (2007). Pre-mRNA 3' cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. RNA Biol 4, 26-33.

19. Balbo, P. B., Meinke, G. & Bohm, A. (2005). Kinetic studies of yeast polyA polymerase indicate an induced fit mechanism for nucleotide specificity. Biochemistry 44, 7777-7786.

20. Edmonds, M. (2002). A history of poly A sequences: from formation to factors to function. Prog Nucleic Acid Res Mol Biol 71, 285-389.

21. Meyer, S., Urbanke, C. & Wahle, E. (2002). Equilibrium studies on the association of the nuclear poly(A) binding protein with poly(A) of different lengths. Biochemistry 41, 6082-6089.

22. Keller, R. W., Kuhn, U., Aragon, M., Bornikova, L., Wahle, E. & Bear, D. G. (2000). The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail. J Mol Biol 297, 569-583.

23. Hirose, Y. & Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. Nature 395, 93-96.

24. Licatalosi, D. D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J. B. & Bentley, D. L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Mol Cell 9, 1101-1111.

25. Ryan, K., Murthy, K. G., Kaneko, S. & Manley, J. L. (2002). Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3' cleavage. Mol Cell Biol 22, 1684-1692.

26. Takagaki, Y. & Manley, J. L. (2000). Complex protein interactions within the human polyadenylation machinery identify a novel component. Mol Cell Biol 20, 1515-1525.

27. Zhao, J., Kessler, M., Helmling, S., O'Connor, J. P. & Moore, C. (1999). Pta1, a component of yeast CF II, is required for both cleavage and poly(A) addition of mRNA precursor. Mol Cell Biol 19, 7733-7740.

28. Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates III, J. R., Frank, J. & Manley, J. L. (2009). Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. Mol Cell 33, 365-376.

29. Keller, W. & Minvielle-Sebastia, L. (1997). A comparison of mammalian and yeast pre-mRNA 3'-end processing. Curr Opin Cell Biol 9, 329-336.

30. He, X., Khan, A. U., Cheng, H., Pappas, D. L., Jr., Hampsey, M. & Moore, C. L. (2003). Functional interactions between the transcription and mRNA 3' end processing machineries mediated by Ssu72 and Sub1. Genes Dev 17, 1030-1042.

31. Dichtl, B., Aasland, R. & Keller, W. (2004). Functions for S. cerevisiae Swd2p in 3' end formation of specific mRNAs and snoRNAs and global histone 3 lysine 4 methylation. Rna 10, 965-977.

32. Skaar, D. A. & Greenleaf, A. L. (2002). The RNA polymerase II CTD kinase CTDK-I affects pre-mRNA 3' cleavage/polyadenylation through the processing component Pti1p. Mol Cell 10, 1429-1439.

33. He, X. & Moore, C. (2005). Regulation of yeast mRNA 3' end processing by phosphorylation. Mol Cell 19, 619-629.

34. Hofmann, I., Schnolzer, M., Kaufmann, I. & Franke, W. W. (2002). Symplekin, a constitutive protein of karyo- and cytoplasmic particles involved in mRNA biogenesis in Xenopus laevis oocytes. Mol Biol Cell 13, 1665-1676.

35. Barnard, D. C., Ryan, K., Manley, J. L. & Richter, J. D. (2004). Symplekin and xGLD-2 are required for CPEB-mediated cytoplasmic polyadenylation. Cell 119, 641-651.

36. Kolev, N. G. & Steitz, J. A. (2005). Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs. Genes Dev 19, 2583-2592.

37. Vethantham, V., Rao, N. & Manley, J. L. (2007). Sumoylation modulates the assembly and activity of the pre-mRNA 3' processing complex. Mol Cell Biol 27, 8848-8858.

38. Mitic, L. L. & Anderson, J. M. (1998). Molecular architecture of tight junctions. Annu Rev Physiol 60, 121-142.

39. Langbein, L., Pape, U. F., Grund, C., Kuhn, C., Praetzel, S., Moll, I., Moll, R. & Franke, W. W. (2003). Tight junction-related structures in the absence of a lumen: occludin, claudins and tight junction plaque proteins in densely packed cell formations of stratified epithelia and squamous cell carcinomas. Eur J Cell Biol 82, 385-400.

40. Keon, B. H., Schafer, S., Kuhn, C., Grund, C. & Franke, W. W. (1996). Symplekin, a novel type of tight junction plaque protein. J Cell Biol 134, 1003-1018.

41. Kiessling, F., Kartenbeck, J. & Haller, C. (1999). Cell-cell contacts in the human cell line ECV304 exhibit both endothelial and epithelial characteristics. Cell Tissue Res 297, 131-140.

42. Paffenholz, R., Kuhn, C., Grund, C., Stehr, S. & Franke, W. W. (1999). The arm-repeat protein NPRAP (neurojungin) is a constituent of the plaques of the outer limiting zone in the retina, defining a novel type of adhering junction. Exp Cell Res 250, 452-464.

43. Kavanagh, E., Buchert, M., Tsapara, A., Choquet, A., Balda, M. S., Hollande, F. & Matter, K. (2006). Functional interaction between the ZO-1-interacting transcription factor ZONAB/DbpA and the RNA processing factor symplekin. J Cell Sci 119, 5098-5105.

44. Xing, H., Mayhew, C. N., Cullen, K. E., Park-Sarge, O. K. & Sarge, K. D. (2004). HSF1 modulation of Hsp70 mRNA polyadenylation via interaction with symplekin. J Biol Chem 279, 10551-10555.

45. Ghazy, M., He, X., Singh, B. N., Hampsey, M. & Moore, C. (2009). The essential N-terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing. Mol Cell Biol.

46. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

47. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. Bioinformatics 14, 892-893.

48. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299, 499-520.

49. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2006). Pfam: clans, web tools and services. Nucleic Acids Res 34, D247-251.

50. Zdobnov, E. M. & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848.

51. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31, 3635-3641.

52. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein server. Nucleic Acids Res 32, W321-326.

53. Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369-3376.

54. Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. Science 252, 1162-1164.

55. Stols, L., Gu, M., Dieckman, L., Raffen, R., Collart, F. R. & Donnelly, M. I. (2002). A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. Protein Expr Purif 25, 8-15.

56. Donnelly, M. I., Zhou, M., Millard, C. S., Clancy, S., Stols, L., Eschenfeldt, W. H., Collart, F. R. & Joachimiak, A. (2006). An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. Protein Expr Purif 47, 446-454.

57. Otwinowski, Z. a. M., W. (1997). *Processing of X-ray Diffraction Data Collected in Oscillation Mode, in Methods in Enzymology, Macromolecular Crystallography, Part A* (Carter, C., Jr., and Sweet, R., Ed.), 276, Academic Press, New York.

58. Fu, Z. Q., Rose, J. & Wang, B. C. (2005). SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination. Acta Crystallogr D Biol Crystallogr 61, 951-959.

59. Sheldrick, G. M. (2008). A short history of SHELX. Acta Crystallogr A 64, 112-122.

60. Terwilliger, T. C. (2003). SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol 374, 22-37.

61. Emsley, P. a. C., Kevin. (2004). Coot: Model-Building Tools for Molecular Graphics. Acta Crystallographica Section D - Biological Crystallography 60, 2126-2132.

62. Collaborative computational project. (1994). The CCP4 Suite: Programs for Protein Crystallography. Acta Cryst. D50, 760-763.

63. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 50, 437-450.

64. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. J Mol Biol 309, 1-18.

65. Chook, Y. M. & Blobel, G. (1999). Structure of the nuclear transport complex karyopherin-beta2-Ran x GppNHp. Nature 399, 230-237.

66. Goldenberg, S. J., Cascio, T. C., Shumway, S. D., Garbutt, K. C., Liu, J., Xiong, Y. & Zheng, N. (2004). Structure of the Cand1-Cul1-Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases. Cell 119, 517-528.

67. Ritco-Vonsovici, M., Ababou, A. & Horton, M. (2007). Molecular plasticity of beta-catenin: new insights from single-molecule measurements and MD simulation. Protein Sci 16, 1984-1998.

68. Sagermann, M., Stevens, T. H. & Matthews, B. W. (2001). Crystal structure of the regulatory subunit H of the V-type ATPase of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 98, 7134-7139.

69. Sampietro, J., Dahlberg, C. L., Cho, U. S., Hinds, T. R., Kimelman, D. & Xu, W. (2006). Crystal structure of a beta-catenin/BCL9/Tcf4 complex. Mol Cell 24, 293-300.

70. Wei, Z., Zhang, P., Zhou, Z., Cheng, Z., Wan, M. & Gong, W. (2004). Crystal structure of human eIF3k, the first structure of eIF3 subunits. J Biol Chem 279, 34983-34990.

71. Xu, Y., Xing, Y., Chen, Y., Chao, Y., Lin, Z., Fan, E., Yu, J. W., Strack, S., Jeffrey, P. D. & Shi, Y. (2006). Structure of the protein phosphatase 2A holoenzyme. Cell 127, 1239-1251.

72. Zachariae, U. & Grubmuller, H. (2006). A highly strained nuclear conformation of the exportin Cse1p revealed by molecular dynamics simulations. Structure 14, 1469-1478.

73. Zachariae, U. & Grubmuller, H. (2008). Importin-beta: structural and dynamic determinants of a molecular spring. Structure 16, 906-915.

74. Holm, L. & Sander, C. (1996). Mapping the protein universe. Science 273, 595-603.

75. Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell 110, 501-512.

76. Neuwald, A. F. & Hirano, T. (2000). HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. Genome Res 10, 1445-1452.

77. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. Structure 8, 329-338.

78. Lee, S. J., Matsuura, Y., Liu, S. M. & Stewart, M. (2005). Structural basis for nuclear import complex dissociation by RanGTP. Nature 435, 693-696.

79. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics 16, 566-567.

80. Fang, X., Chen, T., Tran, K. & Parker, C. S. (2001). Developmental regulation of the heat shock response by nuclear transport factor karyopherin-alpha3. Development 128, 3349-3358.

81. Isgro, T. A. & Schulten, K. (2005). Binding dynamics of isolated nucleoporin repeat regions to importin-beta. Structure 13, 1869-1879.

82. Bai, Y., Auperin, T. C., Chou, C. Y., Chang, G. G., Manley, J. L. & Tong, L. (2007). Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. Mol Cell 25, 863-875.

83. Qu, X., Perez-Canadillas, J. M., Agrawal, S., De Baecke, J., Cheng, H., Varani, G. & Moore, C. (2007). The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3'-end processing. J Biol Chem 282, 2101-2115.

84. Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W. & Doublie, S. (2008). Crystal structure of the 25 kDa subunit of human cleavage factor Im. Nucleic Acids Res 36, 3474-3483.

85. Legrand, P., Pinaud, N., Minvielle-Sebastia, L. & Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. Nucleic Acids Res 35, 4515-4522.

86. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. & Pluckthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J Mol Biol 376, 1282-1304.

87. Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. (2005). The Amber biomolecular simulation programs. J Comput Chem 26, 1668-1688.

88. Jorgensen, W., Chandrasekhar J, Madura JD, Impey RW, Klein ML. (1983). Comparison of simple potential functions for simulating liquid water. J Chem Phys 79, 926-935.

89. Essman U, P. L., Berkowitz ML, Darden T, Lee H, Pedersen L. (1995). A smooth particl mesh Ewald method. J Chem Phys 103, 8577-8593.

90. Teotico, D. G., Frazier, M. L., Ding, F., Dokholyan, N. V., Temple, B. R. & Redinbo, M. R. (2008). Active nuclear receptors exhibit highly correlated AF-2 domain motions. PLoS Comput Biol 4, e1000111.

91. Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X. & Chen, Y. Z. (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31, 3692-3697.

92. LeTilly, V. & Royer, C. A. (1993). Fluorescence anisotropy assays implicate protein-protein interactions in regulating trp repressor DNA binding. Biochemistry 32, 7753-7758.

93. Aviv, T., Lin, Z., Lau, S., Rendl, L. M., Sicheri, F. & Smibert, C. A. (2003). The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. Nat Struct Biol 10, 614-621.

94. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3, 722-732.

95. Lanka, E. & Wilkins, B. M. (1995). DNA processing reactions in bacterial conjugation. Annu Rev Biochem 64, 141-169.

96. Lessl, M. & Lanka, E. (1994). Common mechanisms in bacterial conjugation and Ti-mediated T-DNA transfer to plant cells. Cell 77, 321-324.

97. Zupan, J. R. & Zambryski, P. (1995). Transfer of T-DNA from Agrobacterium to the plant cell. Plant Physiol 107, 1041-1047.

98. Mazel, D. & Davies, J. (1999). Antibiotic resistance in microbes. Cell Mol Life Sci 56, 742-754.

99. de la Cruz, F. & Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol 8, 128-133.

100. Di Laurenzio, L., Frost, L. S. & Paranchych, W. (1992). The TraM protein of the conjugative plasmid F binds to the origin of transfer of the F and ColE1 plasmids. Mol Microbiol 6, 2951-2959.

101. Fekete, R. A. & Frost, L. S. (2002). Characterizing the DNA contacts and cooperative binding of F plasmid TraM to its cognate sites at oriT. J Biol Chem 277, 16705-16711.

102. Lahue, E. E. & Matson, S. W. (1990). Purified Escherichia coli F-factor TraY protein binds oriT. J Bacteriol 172, 1385-1391.

103. Lum, P. L., Rodgers, M. E. & Schildbach, J. F. (2002). TraY DNA recognition of its two F factor binding sites. J Mol Biol 321, 563-578.

104. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell 87, 1295-1306.

105. Howard, M. T., Nelson, W. C. & Matson, S. W. (1995). Stepwise assembly of a relaxosome at the F plasmid origin of transfer. J Biol Chem 270, 28381-28386.

106. Byrd, D. R., Sampson, J. K., Ragonese, H. M. & Matson, S. W. (2002). Structure-function analysis of Escherichia coli DNA helicase I reveals non-overlapping transesterase and helicase domains. J Biol Chem 277, 42645-42653.

107. Stern, J. C. & Schildbach, J. F. (2001). DNA recognition by F factor TraI36: highly sequence-specific binding of single-stranded DNA. Biochemistry 40, 11586-11595.

108. Datta, S., Larkin, C. & Schildbach, J. F. (2003). Structural insights into single-stranded DNA binding and cleavage by F factor TraI. Structure (Camb) 11, 1369-1379.

109. Lujan, S. A., Guogas, L. M., Ragonese, H., Matson, S. W. & Redinbo, M. R. (2007). Disrupting antibiotic resistance propagation by inhibiting the conjugative DNA relaxase. Proc Natl Acad Sci U S A.

110. Lahue, E. E. & Matson, S. W. (1988). Escherichia coli DNA helicase I catalyzes a unidirectional and highly processive unwinding reaction. J Biol Chem 263, 3208-3215.

111. Matson, S. W. & Ragonese, H. (2005). The F-plasmid TraI protein contains three functional domains required for conjugative DNA strand transfer. J Bacteriol 187, 697-706.

112. Ragonese, H., Haisch, D., Villareal, E., Choi, J. H. & Matson, S. W. (2007). The F plasmid-encoded TraM protein stimulates relaxosome-mediated cleavage at oriT through an interaction with TraI. Mol Microbiol 63, 1173-1184.

113. Guogas, L. M., Kennedy, S. A., Lee, J. H. & Redinbo, M. R. (2008). A Novel Fold in the TraI Relaxase-Helicase C-Terminal Domain Is Essential for Conjugative DNA Transfer. J Mol Biol.

114. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25, 4876-4882.

115. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzymology 276, 307-326.

116. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr D Biol Crystallogr 59, 2023-2030.

117. Abrahams, J. P. & Leslie, A. G. (1996). Methods used in the structure determination of bovine mitochondrial F-1 ATPase. Acta Cryst. D52, 30-42.

118. Cowtan, K. & Main, P. (1998). Miscellaneous algorithms for density modification. Acta Crystallogr D Biol Crystallogr 54, 487-493.

119. Emsley, P., Cowtan, K. (2004). Coot: Model-Building Tools for Molecular Graphics. Acta Crystallographica Section D - Biological Crystallography 60, 2126-2132.

120. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53, 240-255.

121. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54 (Pt 5), 905-921.

122. Mills, K. V. & Perler, F. B. (2005). The mechanism of intein-mediated protein splicing: variations on a theme. Protein Pept Lett 12, 751-755.

123. Stephenson, R. C. & Clarke, S. (1989). Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins. J Biol Chem 264, 6164-6170.

124. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J Mol Biol 233, 123-138.

125. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60, 2256-2268.

126. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002). Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 317, 41-72.

127. Ramakrishnan, C., Dani, V. S. & Ramasarma, T. (2002). A conformational analysis of Walker motif A [GXXXXGKT (S)] in nucleotide-binding and other proteins. Protein Eng 15, 783-798.

128. Harris, D. A. (1978). The interactions of coupling ATPases with nucleotides. Biochim Biophys Acta 463, 245-273.

129. Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J. & Kelley, L. A. (2008). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins 70, 611-625.

130. Wigle, T. J. & Singleton, S. F. (2007). Directed molecular screening for RecA ATPase inhibitors. Bioorg Med Chem Lett 17, 3249-3253.

131. Gosselin, S., Alhussaini, M., Streiff, M. B., Takabayashi, K. & Palcic, M. M. (1994). A continuous spectrophotometric assay for glycosyltransferases. Anal Biochem 220, 92-97.

132. O'Rand M, G., Widgren, E. E., Sivashanmugam, P., Richardson, R. T., Hall, S. H., French, F. S., VandeVoort, C. A., Ramachandra, S. G., Ramesh, V. & Jagannadha Rao, A. (2004). Reversible immunocontraception in male monkeys immunized with eppin. Science 306, 1189-1190.

133. O'Rand, M. G., Widgren, E. E., Wang, Z. & Richardson, R. T. (2006). Eppin: an effective target for male contraception. Mol Cell Endocrinol 250, 157-162.

134. Wang, Z., Widgren, E. E., Sivashanmugam, P., O'Rand, M. G. & Richardson, R. T. (2005). Association of eppin with semenogelin on human spermatozoa. Biol Reprod 72, 1064-1070.